Unsupervised Physics Informed Decomposition of Incomplete Time-Resolved Spectroscopy

Tom Muir[1] **Qifeng Liu**[1] **Mohammadrahim Kazemzadeh**[2*] **William Mills**[2]
Aleš Leonardis[1] **Huabing Yin**[2] **Alexander Krull**[1*]

[1]School of Computer Science, University of Birmingham
[2]School of Engineering, University of Glasgow

Abstract

Raman spectroscopy is often hindered by strong autofluorescence backgrounds and noise. While methods for computational removal exist, better results are often achieved through a time-intensive process of waiting for fluorescence to bleach before recording the Raman spectrum. Here, we instead show that short, incomplete recordings of the bleaching process already contain enough information to recover the underlying Raman signal. We present the first unsupervised model that operates on short, incomplete, variable-length sequences of time-resolved measurements, decomposing them into Raman spectra and autofluorescence components while simultaneously removing measurement noise. Our VAE-based method simultaneously learns (i) a bank of fluorophore spectra comprising the baseline, (ii) possible peak positions and shapes, (iii) a model of the measurement noise, and (iv) neural networks for the decomposition of Raman spectra and decay behaviour. We evaluate our method on spectral reconstruction and downstream peak detection, where it outperforms state-of-the-art approaches on multiple simulated benchmarks as well as on a newly collected real-world dataset of spectral time series. We further show that normally disregarded fluorescence decay dynamics themselves may contain discriminative information that can be exploited in future work.

1 Introduction

Raman scattering [1] is the inelastic scattering of light by molecular vibrations, producing wavelength shifts characteristic of the chemical composition of a sample. Raman spectroscopy [2] is an essential tool in chemistry, geology, and life science, enabling non-destructive analysis of chemical bonds and widely used to analyse cells and materials or diagnose disease. Unfortunately, the Raman effect is inherently inefficient, meaning only a tiny fraction of light is affected. This makes Raman spectroscopy highly susceptible to measurement noise. Furthermore, Raman spectroscopy suffers from strong autofluorescence baselines overlaid on the spectrum. Baselines are often many times stronger than the original spectrum and further exacerbate shot-noise contamination. In practice, fluorescence baselines can be reduced by photobleaching under the excitation laser [3], either through deliberate pre-exposure or by repeated acquisition until the baseline diminishes. However, complete photobleaching is time-consuming, and several computational methods have therefore been proposed to remove baselines and noise from single spectra algorithmically or through supervised learning. Such methods struggle when the data are very noisy or when the test-data distribution deviates from the simulated training data.

Another avenue, suggested very recently [4], is to analytically separate the baseline from a bleaching time series by least squares, treating it as a single fluorescence spectrum whose intensity decreases

*Corresponding authors: a.f.f.krull@bham.ac.uk, mohammadrahim.kazemzadeh@glasgow.ac.uk

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

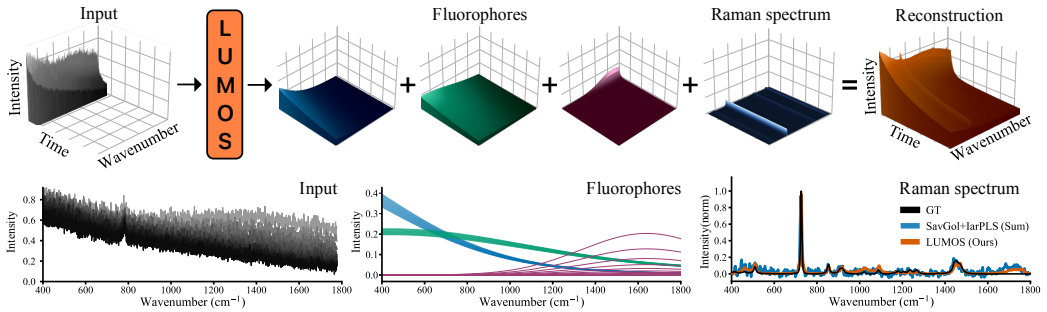


Figure 1: Spectral Raman time series consist of consecutive spectral measurements, in which the true Raman spectrum is obscured by a diminishing fluorescent baseline. **Top:** LUMOS takes a partially observed spectral time series as input and decomposes it into a set of exponentially decaying fluorophores and a constant Raman spectrum. LUMOS outperforms state-of-the-art methods in the extraction of Raman spectra on real and multiple simulated datasets, while additionally recovering fluorophore decay behaviour, which may itself carry valuable information. **Bottom:** Superimposed input spectra of the time series, the predicted fluorophores, and the extracted Raman spectrum compared with ground truth and a classical baseline, for an example from the Pharmaceutical Ingredients dataset [8].

over time. However, such methods fail to account for the complexity of real-life bleaching, where the baseline is composed of multiple fluorophores, each decaying at its own rate. In a biological setting, this can even mean that the same fluorophore decays at variable rates depending on its microenvironment within the cell [5, 6].

Here, we present LUMOS (Latent Un-Mixing of partially Observed Spectra), a novel physics-informed machine learning approach that combines a variational autoencoder (VAE) [7] with an analytical model of fluorescence decay and measurement noise. Our model can decompose noisy incomplete sequences into an estimated Raman spectrum and multiple fluorophores with their decay profiles (Figure 1). During training, our model learns a bank of fluorophore spectra, the locations and shapes of potential Raman peaks, and a parametric model of measurement noise. Together with an encoder and decoder network, this model describes the distribution of Raman spectra, fluorophore abundances, and decay behaviours. This physics-informed approach provides a strong inductive bias, making the method highly data-efficient and enabling training from as few as 240 measurements. Our approach is fully unsupervised: it never observes clean Raman spectra, and its physical priors constrain the space of solutions rather than providing labels. A single configuration is used for all datasets, without per-dataset hyperparameter tuning.

We evaluate our model on four simulated datasets and one newly introduced dataset of 431 measured Raman bleaching time series. Our method outperforms all non-learning and unsupervised approaches, including state-of-the-art machine learning methods, in recovering Raman spectra, and in most cases also outperforms a supervised method. Additionally, we test the downstream behaviour of our model output for peak detection and again find that we are able to outperform baseline methods in many cases. In our ablation study, we explore the sensitivity of our method to hyperparameter choice, sequence length, and measurement noise. Finally, we investigate the properties of the fluorophore decay behaviour that LUMOS extracts from our new dataset and find indications that fluorescence decay may contain meaningful information beyond the Raman spectrum. In summary, our contributions are:

- We introduce LUMOS, a physics-informed VAE for unsupervised decomposition of incomplete Raman bleaching time series, available at `https://github.com/krulllab/LUMOS`.
- We release a public dataset of human cell Raman spectral time series, accessible via the same repository: `https://github.com/krulllab/LUMOS#data`.
- We demonstrate that LUMOS improves spectral quality and peak detection and provide an ablation study analysing its sensitivity to hyperparameters, sequence length, and noise.
- We analyse the extracted fluorophore decay behaviour and find evidence that bleaching dynamics may contain meaningful sample information beyond the recovered Raman spectrum.

2 Related Work

2.1 Non-learning Methods

Several traditional algorithmic methods have been proposed for baseline removal in Raman spectroscopy. AirPLS [9] is an iterative reweighted least-squares approach that fits a baseline by minimising a reweighted quadratic error with respect to the input, while simultaneously enforcing baseline smoothness through an additional regularisation term. This method was improved in [10] by introducing a statistically motivated weighting function. In [11], it was further improved (IarPLS) by modifying the weighting scheme and objective function. Although AirPLS and its variants are widely used today, a major drawback is that a user must manually set several parameters. Another approach, IModPoly [12], enforces smoothness by robustly fitting a low-order polynomial. An older method, IModPoly still achieves respectable results while requiring fewer user-adjusted parameters.

All of the above methods are designed to handle baselines, but not the measurement noise that can additionally affect recorded spectra. The de facto standard approach for handling measurement noise in the field is the Savitzky-Golay (SG) method [13]. SG works by sliding a window over the spectrum, locally fitting a polynomial to the data, and then retrieving the value of the central point of the window from the polynomial. Traditionally, SG is often used jointly with methods such as AirPLS, IModPoly, or related approaches [14]. While these methods are related to our work in that they address noise and baseline removal, they rely purely on heuristics such as smoothness or locality, which must be controlled through parameters and cannot be learned from data.

Very recently, Shiozaki et al. [4] suggested an algorithmic approach for using spectral bleaching time series to extract the Raman spectrum. However, in contrast to our method, the authors rely on the assumption that the baseline corresponds to a single fluorophore, require long acquisitions, and do not remove measurement noise. Unfortunately, Shiozaki et al. do not provide public code for their approach.

2.2 Supervised Methods

Recent years have seen the introduction of several supervised learning-based methods. In [15], the authors train a UNet [16] on simulated data to denoise Raman spectra and remove baselines. Similarly, [17] use a UNet variant with simulated noisy data, albeit limited purely to the denoising task. In [18], the authors present a two-stage deep learning pipeline that partially operates in the Fourier domain and is combined with polynomial baseline subtraction. They focus on designing a highly realistic simulator, which is used to generate training data. Similar to our method, they use domain knowledge to design their model. However, unlike [18], our method leverages measurement physics to enable unsupervised training.

In [19], Tang et al. present a supervised method for processing Raman images that couples a spectral module with an image module. The problem is similar in spirit to ours, as it uses additional spatial input dimensions from hyperspectral imaging, whereas we exploit temporal bleaching dynamics. A different approach is suggested by [20]. The authors train a model to predict the parameters of an iterative algorithmic approach, thereby mitigating the problem of parameter tuning.

2.3 Self- and Unsupervised Methods

In [21], Wu et al. describe a self-supervised denoising method based on splitting noisy spectra into non-overlapping subsampled versions and predicting one from the other. In [22], Bench et al. use a CycleGAN [23] to train a denoiser for Raman spectroscopy based on unpaired clean and noisy data. However, these methods do not support baseline removal. The RSPSSL [24] approach provides a more comprehensive self-supervised method capable of Raman preprocessing, including denoising and baseline subtraction. RSPSSL uses a combination of two networks: a generative adversarial network (GAN) [25] generates realistic paired training data for a second supervised network, which then processes the spectra. Unfortunately, the software is closed-source, and the authors only provide a web-based demo that does not allow for the processing of whole datasets. Georgiev et al. [26] present a method that is similar in spirit to our approach. The paper introduces an autoencoder-based, physics-informed unsupervised method for decomposing Raman spectra. However, unlike our method, it uses hyperspectral data as input instead of time series. Moreover, while [26] is fully deterministic, we adopt a Bayesian perspective based on a VAE architecture.

In our evaluation, we focus on Sjöberg et al. [27], who recently presented a method for Raman denoising and baseline removal with a publicly available implementation. While [27] is trained in a supervised manner on simulated data, the authors claim versatility across different datasets and suggest direct applicability of their published pre-trained model. Here, we treat this pretrained version as an unsupervised method, but additionally fine-tune it on each dataset to derive a supervised variant.

3 Method

3.1 Formation of Spectral Time Series

The Raman Spectrum of a sample can be described as a vector $\boldsymbol{s} = (s_1, \ldots, s_N)$, with each component $s_{\tilde{\nu}}$ capturing the intensity at wavenumber $\tilde{\nu}$. Raman spectra are composed of individual peaks, corresponding to the vibrational modes of molecules in the sample. The shape of these peaks can be well described as the Voigt profile [28]: a convolution of a Lorentzian and Gaussian shape. The Raman spectrum is a weighted sum of Voigt profiles, with

$$s_{\tilde{\nu}} = \sum_{j=1}^{m} \alpha^j V_{\tilde{\nu}}^j, \tag{1}$$

where $s_{\tilde{\nu}}$ is the intensity at wavenumber $\tilde{\nu}$ and $V_{\tilde{\nu}}^j$ is the j-th Voigt profile, parameterised by a centre μ^j, a scale γ^j for the Lorentzian component, and a shared Gaussian width σ_G which accounts for the instrument response and is shared among all peaks. We use $\alpha^j \geq 0$ to denote the mixing coefficients describing the strength of each peak. Unfortunately, when attempting to measure a spectrum $\boldsymbol{s}$, the true signal is obscured by a fluorescent baseline discussed below.

Fluorescent Baselines change over time due to photobleaching, which means they slowly decay when multiple spectra are recorded in sequence. In each acquisition a CCD detector integrates light over an exposure time Δt. Taking this into account, we model the baseline-affected signal at acquisition τ as

$$y_{\tilde{\nu},\tau} = \Delta t \cdot s_{\tilde{\nu}} + F_{\tilde{\nu},\tau}, \tag{2}$$

where $F_{\tilde{\nu},\tau}$ is the baseline for acquisition τ. Generally, samples contain a mixture of fluorophores contributing to this baseline, each with its own abundance and decay rate. We model the baseline as a sum of K exponentially decaying fluorophores. Here, we use t_τ and $t_{\tau+1}$ to denote the beginning and end of the exposure period for measurement τ, with $t_{\tau+1} - t_\tau = \Delta t$. We can thus compute

$$F_{\tilde{\nu},\tau} = \sum_{k=1}^{K} w^k \cdot B_{\tilde{\nu}}^k \int_{t_\tau}^{t_{\tau+1}} e^{-t\lambda^k}\, dt, \tag{3}$$

where $B_{\tilde{\nu}}^k$ is the spectrum of fluorophore k, w^k is its abundance in the sample, and λ^k is its decay rate. As rates are predicted per sample and the baseline comprises several such components, the model also has the potential to represent the multi-exponential bleaching that arises in heterogeneous environments.

Measurement Noise in Raman spectroscopy is often modelled as a combination of Poisson shot noise and Gaussian readout noise [17]. Both types of noise are conditionally independent in time and across wavenumbers given an underlying signal. We can thus compute the probability of measuring a spectral time series $\boldsymbol{x}$ given the noise-free baseline-affected signal $\boldsymbol{y}$ as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{\tilde{\nu}=1,\tau=1}^{N,T} p(x_{\tilde{\nu},\tau}|y_{\tilde{\nu},\tau}), \tag{4}$$

where $p(x_{\tilde{\nu},\tau}|y_{\tilde{\nu},\tau})$ is the combination of Poisson and Gaussian noise for a single wavenumber and measurement time. Here, we use a simple approximation [29] assuming $p(x_{\tilde{\nu},\tau}|y_{\tilde{\nu},\tau})$ follows a normal distribution centred at $y_{\tilde{\nu},\tau}$ with signal-dependent (heteroscedastic) variance

$$\sigma_{\tilde{\nu},\tau}^2 = a \cdot y_{\tilde{\nu},\tau} + b, \tag{5}$$

where a and b are constants that depend on the relative intensity of the Gaussian readout noise, as well as on scaling and shifting of the data during digitisation.

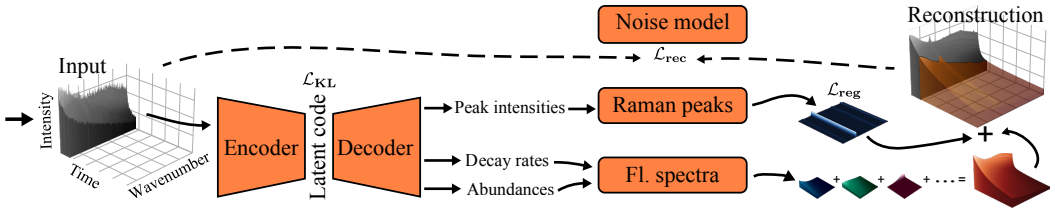


Figure 2: LUMOS is trained as a VAE. A partially observed noisy spectrum is encoded into a latent space posterior, where the KL loss is computed. The decoder takes a latent sample and produces peak intensities for each of the globally learned Raman peaks, as well as decay rates and abundances for each of the globally learned fluorophore spectra. The reconstruction is computed as the sum of decaying fluorophores and the constant Raman spectrum. We use an ℓ_1 regularisation loss on the Raman spectrum. Finally, our globally learned noise model is used to compute the reconstruction loss between input and the corresponding part (grey) of the reconstruction.

3.2 Variational Autoencoder Model

We follow the modelling approach from [30] for training VAEs from noisy data and extend it using our physics-informed model. An overview can be found in Figure 2. Our VAE consists of a fixed prior distribution $p(\boldsymbol{z})$ over a latent space, an encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, and a decoder model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. Following [7], we set the prior as a standard normal distribution over the latent dimensions.

Our Encoder network implements $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ as a 2D convolutional neural network (CNN) taking an observed spectral time series $\boldsymbol{x}$ as input and predicting the parameters (mean and variance) of a normal distribution for each latent dimension. However, unlike images, spectral time series are not translation invariant with respect to the wavenumber dimension. That is, while in an image the same structure could potentially appear anywhere, this is not the case for the peaks in a Raman spectrum, which have meaningful absolute position. To account for this, we add two extra input channels holding the wavenumber and time coordinates, linearly normalised to $[-1, 1]$, so that the network has access to absolute position in both dimensions. The encoder culminates in an adaptive average pooling layer over the temporal dimension, allowing it to handle varying input sequence length T.

Our Physics-Informed Decoder implements $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ based on a global set of learnable Voigt profiles and fluorophore spectra (see Section 3.1), which are learned alongside the decoder model. We use a multi-layer perceptron taking $\boldsymbol{z}$ as input to predict (i) the intensities $\hat{\alpha}^j$ for each of the Voigt profiles, (ii) the abundances $\hat{w}^k$, and (iii) the decay rates $\hat{\lambda}^k$ of each of the baseline fluorophores. Based on a latent code $\boldsymbol{z}$, our decoder uses Eqs. 1 to 3 to compute a noise-free baseline-affected spectrum $\hat{\boldsymbol{y}}$, allowing us to compute

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = p(\boldsymbol{x}|\hat{\boldsymbol{y}}) = \prod_{\tilde{\nu}=1,\tau=1}^{N,T} p(x_{\tilde{\nu},\tau}|\hat{y}_{\tilde{\nu},\tau}), \tag{6}$$

based on Eq. 4, again following [30].

Our Training Procedure directly follows [7, 30]. However, in our case, we start with a training set of spectral time series $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n\}$. Importantly, these time series can be incomplete in the sense that they still contain a substantial amount of residual unbleached fluorescence baseline in their final measurement, so that an approximately pure Raman spectrum is never observed. During training, we sample spectral time series $\boldsymbol{x}^j$ from our training set, encode them, sample in latent space $\boldsymbol{z}^j \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^j)$ using the reparameterisation trick, and decode the sample using our decoder to compute $\hat{\boldsymbol{y}}$ according to Eqs. 1 to 3. Finally, we compute the standard KL loss $\mathcal{L}_{\mathbf{KL}}$ as well as the reconstruction loss as $\mathcal{L}_{\mathbf{rec}} = -\ln p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) + C$ [30], where C is a constant. Based on Eq. 6 and considering that $p(x_{\tilde{\nu},\tau}|y_{\tilde{\nu},\tau})$ is modelled as a normal distribution, we get

$$\mathcal{L}_{\mathbf{rec}} = \sum_{\tilde{\nu}=1,\tau=1}^{N,T} \frac{1}{2} \ln\left(a\hat{y}_{\tilde{\nu},\tau} + b\right) + \frac{(\hat{y}_{\tilde{\nu},\tau} - x_{\tilde{\nu},\tau})^2}{2(a\hat{y}_{\tilde{\nu},\tau} + b)}, \tag{7}$$

where a and b are the parameters from our noise model in Eq. 5. We combine our reconstruction loss with the standard KL loss $\mathcal{L}_{\mathbf{KL}}$ and a simple ℓ_1 regularisation loss $\mathcal{L}_{\mathbf{reg}}$ computed on the predicted

spectra $\hat{\boldsymbol{s}}$. Our combined loss function is

$$\mathcal{L} = \mathcal{L}_{\mathbf{rec}} + \mathcal{L}_{\mathbf{KL}} + \rho \cdot \mathcal{L}_{\mathbf{reg}}, \tag{8}$$

where ρ is a hyperparameter controlling regularisation. Note that no parameter for the balance between $\mathcal{L}_{\mathbf{rec}}$ and $\mathcal{L}_{\mathbf{KL}}$ is required, as it would be redundant with the learnable noise model, which implicitly provides a balancing factor $\frac{1}{2\sigma^2_{\hat{\nu},\tau}}$ for each pixel in Eq. 7. Optimising Eq. 8 allows us to jointly learn the network parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, the bank of fluorophores (described as Gaussian mixture models), and Voigt profiles and additionally the noise model parameters a and b.

Inference in our model is performed by encoding a spectral sequence and decoding the predicted mean of the distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ directly to obtain a spectrum. This avoids latent sampling entirely and is faster than Monte Carlo averaging [30], whilst still providing a reliable point estimate.

Architectural and Implementation Details of our model and the required compute are provided in Appendix sections C.1 and C.2.

4 Experiments and Results

4.1 Datasets

We evaluate on one real-world time-resolved dataset and four synthetic datasets. Each synthetic dataset (see Appendix section A) is composed of real Raman spectra and fluorophores, combined in a synthetic time series generated using the full forward model of Eqs. 2–4. All data is split 70/10/20 into train, validation and test sets using stratified sampling by class label where possible. As LUMOS requires no ground truth, it can, like self-supervised denoisers [31], be trained on the very data it is to process; we therefore train it on all inputs, including those of the test set, without ever observing their ground-truth spectra. All methods are evaluated using the same test split.

The Human Cell dataset is a newly collected real time-resolved dataset which is publicly available at `https://github.com/krulllab/LUMOS`. It consists of 431 photobleaching sequences acquired from human cell samples across four classes: bone, cartilage, stem cells, and MCF-7 breast cancer cells at Δt=0.1 s per frame. We retain the 266 sequences that are long enough to provide both the input frames and a disjoint set of 16 reference frames, and that pass kinetic quality control (Appendix C.4). Ground-truth Raman spectra are approximated by averaging the final 16 valid frames, followed by baseline correction and Savitzky-Golay smoothing. Full acquisition and preprocessing details are provided in Appendix C.4.

Pharmaceutical Ingredients [8] comprises 3,510 spectra from 32 chemical substances including organic solvents and reagents used in active pharmaceutical ingredient (API) development.

Cancer Cell Metabolites (NH_2) [32] comprises 632 SERS spectra of cancer cell metabolites collected on gold nanourchins functionalised with the NH_2 moiety, across 12 cell line classes targeting proteins and lipids for cell line identification.

RRUFF Minerals [33] is a comprehensive collection of Raman spectra from over 1,000 mineral species, representing diverse crystallographic structures and chemical compositions measured under varying experimental conditions including 532 nm and 785 nm excitation.

Mutant Wheat [34] contains Raman spectra from the seventh generation of salt-stress-tolerant wheat mutant lines and their commercial cultivars, acquired at 785 nm excitation and targeting biochemical shifts in carotenoids and protein-related bands for agricultural phenotyping.

4.2 Baseline Methods

While our method takes a sequence of T frames as input, single-frame baselines operate on one exposure. To provide a fair comparison, we evaluate all baseline methods in two modes: The *sum* mode simulates a long exposure by summing all T frames. In synthetic datasets, we sum the clean images of the time series and then apply a single Poisson-Gaussian noise realisation afterwards in order to correctly model the advantage of single long CCD exposure over T individual readouts; readout noise occurs only once per acquisition and a long exposure is therefore preferable in this respect. On the real Human Cells dataset, where clean frames and long exposures are unavailable, we

use the raw sum of the first T frames directly. We additionally evaluate baselines in *last* mode, using frame T of the input series, which contains the least residual fluorescence.

RADAR is evaluated in two settings: unsupervised, using the pretrained weights provided by the authors without modification, and *fine-tuned*, where the pretrained model is adapted independently on each dataset. Details are provided in Appendix C.3.

iModPoly/IarPLS + Savitzky-Golay first apply Savitzky-Golay denoising and then iModPoly/IarPLS, respectively. iModPoly uses polynomial order 5. IarPLS uses $\lambda{=}10^6$. These were qualitatively determined. As noise is a limiting factor, we do a grid search for the SG window size and polynomial order and pick the values to maximise PSNR (compared to GT) on the first 50 test samples, providing the strongest possible comparison for baselines.

4.3 Experiments

Quality of extracted spectra: We compare the quality of extracted spectra in Table 1 and report PSNR, scale-invariant PSNR (SI-PSNR) [35], Pearson correlation (r), and SAM [36]. We find that LUMOS outperforms all non-learning and unsupervised competitors on every dataset and metric. The supervised versions of RADAR are the strongest competitors; LUMOS outperforms them in most cases, including on every spectral metric on the real Human Cell data, with RRUFF and Mutant Wheat the exceptions. Qualitative results can be found in Figure 3. In many places, LUMOS is able to follow general contour more closely than RADAR, while at the same time containing less high-frequency noise compared to iModPoly+SG. Further qualitative results can be found in Appendix section E.

Table 1: Evaluated Raman spectra. Best super/unsupervised results in **bold**, best overall in *italic*.

Dataset	Metric	Non-Learning and Unsupervised							Supervised	
		LUMOS	IarPLS+SG		iModPoly+SG		RADAR		RADAR	
		16 Fr.	Sum	Last	Sum	Last	Sum	Last	Sum	Last
H. Cells	PSNR ↑	***19.55***	18.66	15.38	15.86	14.70	15.10	18.98	14.76	**18.98**
	SI-PSNR ↑	***23.60***	22.10	19.91	20.40	18.63	19.92	21.53	19.67	**21.56**
	Pearson ↑	***0.94***	0.90	0.82	0.88	0.81	0.81	0.88	0.80	**0.89**
	SAM ↓	***15.85***	19.59	25.59	29.92	34.85	26.81	19.98	27.38	**19.92**
Pharma	PSNR ↑	***32.82***	25.16	20.26	26.33	22.20	28.01	25.12	**31.77**	27.36
	SI-PSNR ↑	***33.84***	27.50	23.31	28.46	23.87	30.18	26.37	**32.42**	28.06
	Pearson ↑	***0.97***	0.86	0.71	0.92	0.76	0.95	0.88	***0.97***	0.92
	SAM ↓	***10.98***	24.22	28.46	23.10	37.39	18.78	26.26	**12.14**	20.03
Cancer	PSNR ↑	***21.88***	15.50	13.97	13.11	12.34	14.02	13.19	**20.81**	18.95
	SI-PSNR ↑	***22.48***	20.00	17.28	19.41	17.07	16.62	16.37	**21.49**	19.61
	Pearson ↑	***0.93***	0.85	0.73	0.84	0.71	0.66	0.62	**0.91**	0.86
	SAM ↓	***13.34***	27.86	35.26	39.16	44.87	34.68	38.69	**15.01**	18.49
RRUFF	PSNR ↑	**31.29**	20.75	18.23	27.30	21.97	29.38	25.26	***34.21***	28.24
	SI-PSNR ↑	**31.98**	24.57	22.24	28.17	22.97	30.51	26.22	***34.90***	29.10
	Pearson ↑	**0.96**	0.75	0.67	0.91	0.74	0.92	0.85	***0.98***	0.93
	SAM ↓	**13.58**	34.15	42.88	21.25	38.11	17.91	26.74	***9.26***	18.03
Wheat	PSNR ↑	**28.41**	25.58	21.25	18.64	16.80	16.39	16.83	***32.36***	27.55
	SI-PSNR ↑	**31.73**	27.42	22.80	23.18	19.97	17.64	18.87	***33.24***	28.67
	Pearson ↑	***0.99***	0.97	0.92	0.94	0.88	0.80	0.83	***0.99***	0.98
	SAM ↓	**6.49**	11.00	17.42	20.40	27.62	30.90	27.90	***5.03***	8.50

Peak detection: To assess the downstream applicability, we apply scipy.signal *find_peaks(...)* to the predicted Raman spectra and compare them to ground truth peak positions via detection metrics. In Table 2, we report average precision (AUPRC), the maximum F1 score, as well as the average absolute positional error (P. err.) in cm^{-1} and the average absolute peak height error (H.err.). We rank detections based on peak prominence and choose a matching threshold of $5\,\mathrm{cm}^{-1}$ between peaks to compute true positives, false positives and false negatives. A visual representation of results across

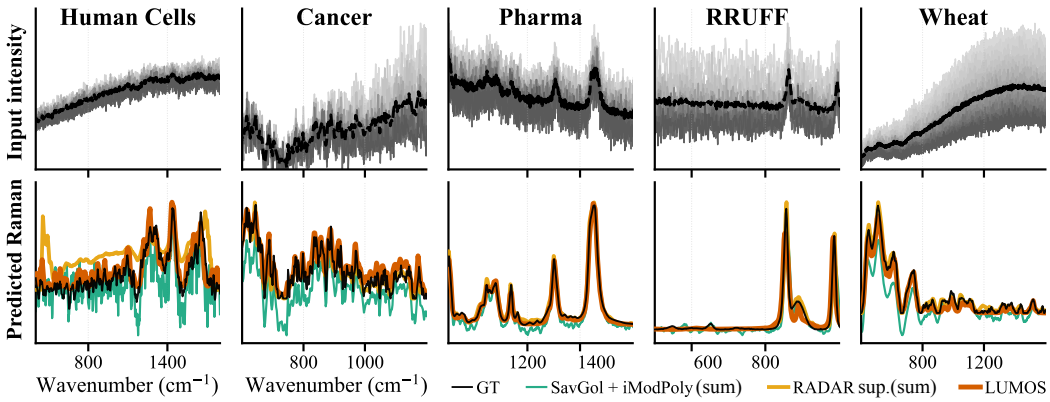


Figure 3: Qualitative comparison of extracted Raman spectra across all four synthetic datasets and the real cell dataset. *Top row:* input frames of the bleaching time series. *Bottom row:* extracted Raman spectra for each method. All methods operate on the first T=16 frames; single-frame baselines receive these as a single long exposure sum. Each column corresponds to a representative test sample from one dataset.

various matching thresholds can be found in Appendix section E. Further evaluation details can be found in Appendix section B.

Table 2: Evaluated peak detection. Best super/unsupervised results in **bold**, best overall in *italic*.

Dataset	Metric	Non-Learning and Unsupervised							Supervised	
		LUMOS	IarPLS+SG		iModPoly+SG		RADAR		RADAR	
		16 Fr.	Sum	Last	Sum	Last	Sum	Last	Sum	Last
H. Cells	AUPRC ↑	***0.38***	0.30	0.19	0.18	0.15	0.22	0.21	**0.21**	**0.21**
	Max F1 ↑	***0.53***	0.46	0.37	0.36	0.33	0.39	0.38	**0.39**	**0.39**
	P. err. ↓	2.42	2.45	2.70	2.60	2.75	***2.41***	2.72	**2.46**	2.70
	H. err. ↓	24.74	27.70	38.24	42.73	46.60	34.30	**22.81**	36.06	***22.14***
Pharma	AUPRC ↑	***0.68***	0.52	0.27	0.55	0.27	0.52	0.31	**0.57**	0.36
	Max F1 ↑	***0.76***	0.60	0.39	0.63	0.39	0.64	0.42	**0.66**	0.47
	P. err. ↓	***1.21***	1.30	2.26	1.35	2.25	1.37	1.96	**1.28**	1.84
	H. err. ↓	***0.03***	0.07	0.17	0.08	0.17	0.08	0.12	**0.04**	0.09
Cancer	AUPRC ↑	**0.58**	0.57	0.23	**0.58**	0.18	0.24	0.20	***0.60***	0.35
	Max F1 ↑	**0.70**	0.69	0.44	**0.70**	0.39	0.43	0.40	***0.71***	0.54
	P. err. ↓	1.38	***1.04***	2.39	1.08	2.40	2.12	2.32	**1.49**	1.88
	H. err. ↓	***0.07***	0.14	0.21	0.21	0.28	0.26	0.29	**0.08**	0.11
RRUFF	AUPRC ↑	**0.40**	0.32	0.20	0.39	0.16	0.39	0.23	***0.46***	0.31
	Max F1 ↑	**0.52**	0.41	0.31	0.49	0.28	0.50	0.33	***0.58***	0.41
	P. err. ↓	2.03	**1.45**	2.27	1.66	2.53	**1.45**	1.96	***1.40***	1.97
	H. err. ↓	**0.08**	**0.08**	0.15	0.10	0.23	0.11	0.18	***0.05***	0.12
Wheat	AUPRC ↑	***0.49***	0.33	0.15	0.31	0.15	0.24	0.12	**0.47**	0.25
	Max F1 ↑	***0.60***	0.49	0.33	0.47	0.33	0.42	0.29	**0.59**	0.38
	P. err. ↓	***1.97***	2.23	2.64	2.53	2.64	2.35	2.62	**2.16**	2.48
	H. err. ↓	**0.04**	0.05	0.09	0.11	0.13	0.12	0.13	***0.02***	0.04

Effects of series length: We vary the number of input frames T and measure reconstruction quality (Figure 4, right). Both methods improve monotonically with T, as longer sequences carry more information about the bleaching dynamics and the Raman spectrum, and LUMOS keeps a consistent advantage over SavGol + iModPoly at every length, with diminishing returns as both saturate. Frame count is proportional to acquisition time, since bleaching is driven by the measurement laser and

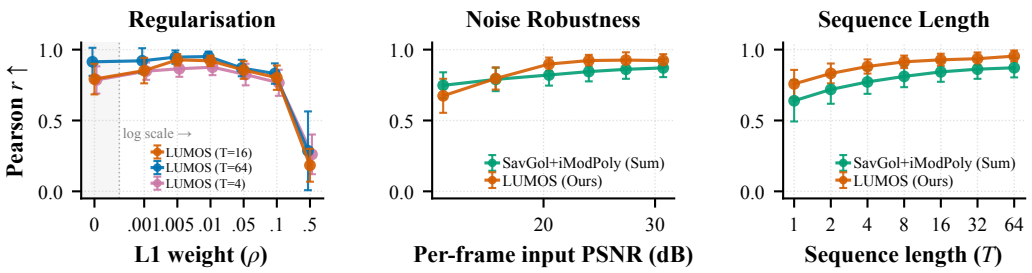


Figure 4: Ablation studies on the Cancer dataset (Pearson r). **Left:** Effect of ℓ_1 regularisation weight ρ. **Center:** Effect of noise level. **Right:** Effect of sequence length T. Error bars indicate standard deviation across test samples.

frames are recorded back to back. With a single frame (0.1 s), where no temporal information is available, LUMOS already exceeds SavGol + iModPoly on 64 frames (6.4 s) in PSNR (16.6 vs. 13.4 dB) and SAM (24.4° vs. 37.3°), showing the strength of its inductive bias even without bleaching dynamics. With 8 frames, it exceeds both baselines on 64 frames on every spectral quality metric (Pearson 0.91 vs. 0.87 and 0.89; Table 3), an 8-fold reduction in acquisition time and laser exposure.

Effects of regularisation: We investigate the sensitivity with respect to the ℓ_1 regularisation hyperparameter ρ and its effect on Pearson correlation (Figure 4, left). We find that regularisation becomes more important for models using shorter time series, while longer sequences seem to carry enough information to make a strong regularisation unnecessary.

Effects of noise: We investigate the effects of imaging noise, by training 6 models on the cancer dataset using 6 intensity levels of Poisson and Gaussian noise (Figure 4, centre). We observe that while performance remains steady for a large range of values, extreme shot noise dominated conditions lead to diminished results. We believe the reason for this could be the simple noise model (Eq. 5), which approximates the noise as a normal distribution with signal-dependent variance.

Number of fluorophores: We investigate the sensitivity with respect to the number of learnable fluorophores K by training models with $K \in \{4, 8, 16, 32\}$, achieving Pearson correlations of 0.64, 0.92, 0.93 and 0.93, respectively (Table 4). Performance degrades for small K and remains stable once K is sufficiently large, including when K exceeds the 16 fluorophores in the library, so $K = 16$ acts as a safe over-specification.

Amortised inference vs. per-sample fitting: Because the VAE relies on amortised inference, we test whether this choice is justified by comparing it with a per-sample fit that uses the same forward model and parameterisation (Table 5). Fitting each sample from scratch achieves a Pearson correlation of only 0.43, and 0.69 even when the true fluorophore spectra are provided. Our model matches even the most informed per-sample variant, in which the true decay rates and abundances are also given and only the Raman peaks are learned (Pearson 0.93 vs. 0.92), and exceeds it in SI-PSNR (22.5 vs. 22.2 dB) and AUPRC (0.58 vs. 0.33). This suggests that the network recovers the baseline accurately, leaving only the high-frequency Raman detail to be modelled, and that it models this detail better than a per-sample fit. Not only do these results demonstrate the benefit of amortised learning via a neural network such as a VAE, but its probabilistic formulation provides the potential for further probabilistic analysis of outputs.

Fluorophore recovery: We compare the learnt fluorophore spectra and their parameters with the ground truth. Since learning is unsupervised, the correspondence between learnt and true fluorophores is unknown, so we match them by Hungarian assignment on the Pearson correlation of their spectra, both for the 16-fluorophore library used throughout and for a smaller, more realistic 4-fluorophore library (Table 6). With 4 fluorophores, learnt spectra closely match the ground truth (r=0.99), and abundances and decay rates are recovered well (r=0.83 and 0.74), on par with a model given the true spectra. With 16 fluorophores, spectra are still matched (r=0.86), but abundances and decay rates are poorly recovered (r=0.30 and 0.25), even given the true spectra, while Raman recovery remains accurate (Pearson 0.93).

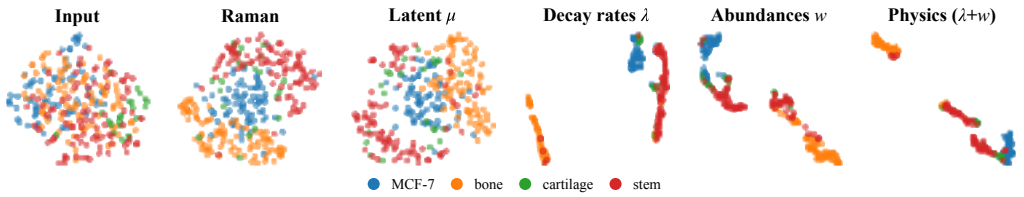


Figure 5: UMAP applied to different extracted properties. Abundances and decay rates lead to substantially different groupings.

Information content of extracted abundances and decay behaviour: To investigate the potential information content of the extracted fluorophore abundances and decay rates, we apply UMAP [37] (Figure 5) to the input, the extracted Raman spectra, the latent means, decay rates λ, abundances w, as well as their combination. We can clearly see that the different cell types are separated in different ways. Using w partially groups stem cells with bone, whereas λ alone groups most stem cells together. This is notable: abundances describe the composition of the baseline, which is largely visible in a single frame, whereas decay rates can only be inferred from the bleaching dynamics, information that conventional acquisition discards.

5 Conclusion

We have introduced LUMOS, an unsupervised model for the simultaneous decomposition of partially observed spectral time series. Research laboratories across biology, medicine, chemistry, geology, and other fields routinely use Raman spectroscopy, yet often struggle with fluorescent baselines that are typically suppressed through time-consuming bleaching procedures. We have shown that high-quality spectra suitable for downstream processing can be acquired using only a fraction of the typical acquisition time by recording short bleaching sequences and training a LUMOS model. Furthermore, our results suggest that fluorescent baselines and their decay behaviour can themselves carry meaningful information, a possibility that remains to be studied extensively. By substantially shortening acquisition times, LUMOS increases throughput and reduces sample photodamage, which may be particularly valuable for photosensitive or live samples.

Limitations and future work: First, the model can struggle with intense shot noise under extremely low-light conditions (Appendix Figure 6), likely due to the simple approximation of the Poisson-Gaussian noise model; noise models that better capture the skewness of the Poisson distribution are a natural extension. Second, LUMOS does not always outperform other methods with respect to positional error in peak detection (Table 2), which may stem from the simplified Voigt peak model. Third, real-data ground truth is necessarily an estimate: a late-frame average, baseline-corrected and smoothed to remove residual fluorescence and noise. Since this reference is constructed with the same kind of baseline correction and smoothing used by the classical baselines, the evaluation may, if anything, favour those methods. Fourth, our simulations share the per-fluorophore exponential decay assumed by LUMOS; while the strong results on real data, whose bleaching is not constrained to this model, suggest robustness to deviations from it, a controlled study of such mismatch, as well as of generalisation across instruments, remains future work. Finally, VAEs in principle allow sampling from a posterior distribution of possible Raman spectra, but in preliminary experiments the resulting uncertainty appears lower than expected and the model seems overconfident at times, motivating an unsupervised calibration strategy. Together, these directions could make LUMOS a more robust tool for rapid Raman acquisition and for exploiting information that is usually discarded during fluorescence bleaching.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by UK Research and Innovation [grant number UKRI251]; and the Medical Research Council [grant number MR/Z50628X/1]. T.M. is supported by a PhD studentship from the School of Computer Science, University of Birmingham. **Competing interests.** The authors declare no competing interests.

References

[1] C. V. Raman and K. S. Krishnan. A New Type of Secondary Radiation. *Nature*, 121(3048): 501–502, March 1928. ISSN 0028-0836, 1476-4687. doi: 10.1038/121501c0.

[2] John R. Ferraro, Kazuo Nakamoto, and Chris W. Brown. *Introductory Raman Spectroscopy*. Elsevier, 2003. ISBN 978-0-12-254105-6. doi: 10.1016/B978-0-12-254105-6.X5000-8.

[3] Kurtulus Golcuk, Gurjit S. Mandair, Andrew F. Callender, Nadder Sahar, David H. Kohn, and Michael D. Morris. Is photobleaching necessary for Raman imaging of bone tissue using a green laser? *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1758(7):868–873, July 2006. ISSN 00052736. doi: 10.1016/j.bbamem.2006.02.022.

[4] Yusuke Shiozaki and Katsumasa Fujita. Photobleaching-assisted extraction of Raman spectral features from fluorescence-enhanced background variation. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 359:127926, October 2026. ISSN 13861425. doi: 10.1016/j.saa.2026.127926.

[5] Jerker Widengren and Rudolf Rigler. Mechanisms of photobleaching investigated by fluorescence correlation spectroscopy. *Bioimaging*, 4(3):149–157, September 1996. ISSN 09669051, 13616374. doi: 10.1002/1361-6374(199609)4:3<149::AID-BIO5>3.0.CO;2-D.

[6] Hellen C. Ishikawa-Ankerhold, Richard Ankerhold, and Gregor P. C. Drummen. Advanced Fluorescence Microscopy Techniques—FRAP, FLIP, FLAP, FRET and FLIM. *Molecules*, 17 (4):4047–4132, April 2012. ISSN 1420-3049. doi: 10.3390/molecules17044047.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[8] Aaron R. Flanagan and Frank G. Glavin. Open-source Raman spectra of chemical compounds for active pharmaceutical ingredient development. *Scientific Data*, 12(1):498, March 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04848-6.

[9] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[10] Sung-June Baek, Aaron Park, Young-Jin Ahn, and Jaebum Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, 140(1):250–257, 2015.

[11] Jianfeng Ye, Ziyang Tian, Haoyun Wei, and Yan Li. Baseline correction method based on improved asymmetrically reweighted penalized least squares for the Raman spectrum. *Applied Optics*, 59(34):10933–10943, 2020.

[12] Feng Gan, Guihua Ruan, and Jinyuan Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and intelligent laboratory systems*, 82(1-2): 59–65, 2006.

[13] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[14] Jiaheng Cui, Xianyan Chen, and Yiping Zhao. Beyond traditional airPLS: Improved baseline removal in SERS with parameter-focused optimization and prediction. *Analytical Chemistry*, 97(30):16211–16218, 2025.

[15] Medhanie Tesfay Gebrekidan, Christian Knipfer, and Andreas Siegfried Braeuer. Refinement of spectra using a deep neural network: Fully automated removal of noise and background. *Journal of Raman Spectroscopy*, 52(3):723–736, 2021.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Yingjie Zeng, Zi-quan Liu, Xian-guang Fan, and Xin Wang. Modified denoising method of Raman spectra-based deep learning for Raman semi-quantitative analysis and imaging. *Microchemical Journal*, 191:108777, 2023.

[18] Mengkun Chen, Sanidhya D. Tripathi, and James W. Tunnell. Simulation-driven deep learning framework for Raman spectral denoising under fluorescence-dominant conditions, 2025. URL `https://arxiv.org/abs/2512.17852`.

[19] Quan Tang, Jiaqi Hu, Jinna Chen, Chenlong Xue, Junfan Chen, Hong Dang, Dan Lu, Huanhuan Liu, Qizhen Sun, Qiaozhou Xiong, et al. A two-stage Raman imaging denoising algorithm based on deep learning. In *2022 Asia Communications and Photonics Conference (ACP)*, pages 2096–2099. IEEE, 2022.

[20] Prasad D Aradhye, Souparna Mandal, Robert D Gray, and Colin J Campbell. Adaptive physics-aware Raman baseline correction with machine learning predicted parameters. *Analytical Chemistry*, 97(48):26708–26719, 2025.

[21] Siyi Wu, Yumin Zhang, Chang He, Zhewen Luo, Zhou Chen, and Jian Ye. Self-supervised learning for generic Raman spectrum denoising. *Analytical Chemistry*, 96(44):17476–17485, 2024.

[22] Ciaran Bench, Mads S. Bergholt, and Mohamed Ali al Badri. Unsupervised denoising of Raman spectra with cycle-consistent generative adversarial networks, 2023. URL `https://arxiv.org/abs/2307.00513`.

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[24] Jiaqi Hu, Gina Jinna Chen, Chenlong Xue, Pei Liang, Yanqun Xiang, Chuanlun Zhang, Xiaokeng Chi, Guoying Liu, Yanfang Ye, Dongyu Cui, et al. RSPSSL: A novel high-fidelity Raman spectral preprocessing scheme to enhance biomedical applications and chemical resolution visualization. *Light: Science & Applications*, 13(1):52, 2024.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[26] Dimitar Georgiev, Álvaro Fernández-Galiana, Simon Vilms Pedersen, Georgios Papadopoulos, Ruoxiao Xie, Molly M Stevens, and Mauricio Barahona. Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders. *Proceedings of the National Academy of Sciences*, 121(45):e2407439121, 2024.

[27] Joel Sjöberg, Nicoleta Siminea, Andrei Păun, Adrian Lita, Mioara Larion, and Ion Petre. RADAR: Raman spectral analysis using deep learning for artifact removal. *Advanced Optical Materials*, 13(25):2500736, 2025.

[28] P. Thompson, D. E. Cox, and J. B. Hastings. Rietveld refinement of Debye–Scherrer synchrotron X-ray data from Al_2 O_3. *Journal of Applied Crystallography*, 20(2):79–83, April 1987. ISSN 0021-8898. doi: 10.1107/S0021889887087090.

[29] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. doi: 10.1109/TIP.2008.2001399.

[30] Mangal Prakash, Alexander Krull, and Florian Jug. Fully unsupervised diversity denoising with convolutional variational autoencoders. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=agHLCOBM5jP`.

[31] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void – learning denoising from single noisy images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2124–2132. IEEE, June 2019. doi: 10.1109/cvpr.2019.00223. URL http://dx.doi.org/10.1109/CVPR.2019.00223.

[32] M. Erzina, A. Trelin, O. Guselnikova, B. Dvorankova, K. Strnadova, A. Perminova, P. Ulbrich, D. Mares, V. Jerabek, R. Elashnikov, V. Svorcik, and O. Lyutakov. Precise cancer detection via the combination of functionalized SERS surfaces and convolutional neural network with independent inputs. *Sensors and Actuators B: Chemical*, 308:127660, April 2020. ISSN 0925-4005. doi: 10.1016/j.snb.2020.127660. URL http://dx.doi.org/10.1016/j.snb.2020.127660.

[33] Barbara Lafuente, Robert T. Downs, Hexiong Yang, and Nate Stone. The power of databases: the RRUFF project. In Thomas Armbruster and Rosa Micaela Danisi, editors, *Highlights in Mineralogical Crystallography*, pages 1–30. De Gruyter, Berlin, Germany, 2015. doi: 10.1515/9783110417104-003.

[34] Ayse Sen, Ibrahim Kecoglu, Muhammad Ahmed, Ugur Parlatan, and Mehmet Burcin Unlu. Differentiation of advanced generation mutant wheat lines: Conventional techniques versus Raman spectroscopy. *Frontiers in Plant Science*, 14, February 2023. ISSN 1664-462X. doi: 10.3389/fpls.2023.1116876. URL http://dx.doi.org/10.3389/fpls.2023.1116876.

[35] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15 (12):1090–1097, 2018.

[36] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image processing system (SIPS) — interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, 44(2–3):145–163, 1993.

[37] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[38] Talley J. Lambert. FPbase: A community-editable fluorescent protein database. *Nature Methods*, 16(4):277–278, April 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0352-8.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

[40] Mario Koddenbrock, Christoph Lange, Robin Legner, Martin Jäger, Martin Kögler, Mariano N Cruz Bournazou, Peter Neubauer, Felix Biessmann, and Erik Rodner. RamanBench: A large-scale benchmark for machine learning on Raman spectroscopy. In *Advances in Neural Information Processing Systems*, 2026. arXiv preprint arXiv:2605.02003.

A Synthetic Data Generation Details

For datasets exceeding 10,000 spectra, we sample a uniform subset of 10,000. We source 16 fluorophore emission spectra from FPbase [38], apply Gaussian smoothing (σ=50 cm^{-1}) to suppress interpolation artefacts, and interpolate them onto the instrument wavenumber grid using PCHIP. Each fluorophore spectrum is L_∞-normalised to unit peak intensity.

For each synthetic sample, 3 fluorophores are drawn without replacement, yielding $\binom{16}{3}$=560 possible combinations.

The 16 fluorophores selected from FPbase are: SYFP2, EYFP, LSSmOrange, mKate2, tdTomato, mOrange2, mRuby3, mApple, DsRed, mVenus, mCherry, vsfGFP-0, DsRed-Express2, TagRFP, mKO2, and mNeonGreen. This selection spans a broad emission range from green to red, covering the spectral diversity of commonly encountered fluorophores and thereby producing complex, overlapping fluorescence backgrounds.

Decay rates are drawn per fluorophore from slow $\lambda \in [0.05, 0.3)$, medium $\lambda \in [0.3, 1.0)$, and fast $\lambda \in [1.0, 5.0]$ s^{-1} components, consistent with multi-timescale photobleaching behaviour of endogenous biological fluorophores [5]. Fluorophore abundances are sampled such that the total fluorescence-to-Raman ratio at t=0 is drawn uniformly from $[3, 15]$, representative of moderate to strong autofluorescence conditions encountered in biological Raman spectroscopy [3]. Raman spectra are sourced from each respective dataset and are L_∞-normalised to unit peak intensity, so that the fluorescence-to-Raman ratio controls the absolute scale of the fluorescence component. Time series are generated using the factored integration model of Eq. 3, and noise is injected per frame following the Poisson-Gaussian model of Eq. 4.

B Peak Detection Evaluation

Peak detection in 1D can be affected by spray-and-pray behaviour, where high AUPRCs can be achieved from very noisy results, which create random detections throughout the spectrum. We observed this issue and found that AUPRCs could be inflated simply by adding noise to the extracted spectra. To address this issue, we use top-K evaluation, where only the best (highest prominence) K detections in each spectrum are considered when computing detection metrics. We set K to the average number of GT peaks per spectrum.

C Technical Details and Hyperparameters

C.1 Implementation for LUMOS and compute details

All models are trained with AdamW [39] with a learning rate of 5×10^{-4}, batch size 8, and gradient clipping at norm 1.0. Training runs for up to 15,000 epochs with 200 steps per epoch, using ReduceLROnPlateau scheduling with patience 20 and early stopping with patience 200. The encoder and decoder use latent dimension $d_{\boldsymbol{z}}$=128 and hidden dimension 512. We use m=300 Voigt peaks with learned positions μ^j, initialised to tile the spectral window evenly, per-peak Lorentzian widths γ^j, initialised to the spacing between peaks, and a shared Gaussian width σ_G, initialised to four times the wavenumber sampling interval. Both FWHMs are constrained to be at least the widest sampling interval of the wavenumber axis, below which a profile cannot be resolved. The fluorophore basis comprises K=16 MoG components each parameterised by C=3 Gaussians. The ℓ_1 regularisation weight is ρ=0.01 and the minimum decay rate is $\lambda_{\min}$=0.001 s^{-1}. All experiments use T=16 input frames at frame duration Δt=0.1 s unless otherwise stated. As our method does not require GT during training, it can be trained on the very data that is to be processed. The model is therefore trained on all available data, including those in the test set, with the validation set reserved for training metrics and early stopping. Our method was trained on an NVIDIA GeForce RTX 3090 GPU with 24 GB GDDR6X memory. Training of each model took approximately 1 hour. Testing required approximately 6.6 ms per sample.

C.2 Model Architecture

Encoder. The encoder is a joint 2D convolutional neural network operating on the input spectrogram $x \in \mathbb{R}^{W \times T}$. We form three input channels by stacking the standardised input signal, clipped at zero, with its wavenumber positional encoding $\nu_{\text{norm}} \in [-1, 1]$, along with a time positional encoding $\tau_{\text{norm}} \in [-1, 1]$. Both of the aforementioned encodings are broadcast to match the $[W, T]$ spatial dimensions. Positional encodings are linearly normalised over the spectral and temporal dimensions respectively, providing the network with explicit coordinate information without the need for learned embeddings.

The convolutional stack comprises four layers with filter sizes of $\{64, 128, 256, 512\}$ and kernel sizes $\{7\times3, 7\times3, 5\times3, 5\times3\}$ (spectral $\times$ temporal). All layers apply a stride of 2 along the spectral axis, and a stride of 1 along the temporal axis, resulting in a $16\times$ reduction, while preserving T. This choice of asymmetric striding allows the network to learn the joint spectro-temporal features, while remaining agnostic to the input length. Further studies could explore the application of models trained on longer time series to shorter time series, leading to the potential realisation of a semi-supervised learning approach.

Following the convolutional stack, the output is reduced to a fixed 16×1 spatial map regardless of the input length T via adaptive average pooling. The aforementioned pooling mechanism makes the encoder length-agnostic: temporal information is aggregated globally, before the fully connected layers. A single fully connected hidden layer of width 512 precedes the μ and $\log \sigma^2$ projection heads, each of dimension $d_z = 128$.

Decoder. The decoder maps a latent sample $z \in \mathbb{R}^{128}$ to three sets of physical parameters via a shared trunk, followed by three independent branches each consisting of a single hidden layer of width 512. The three output heads produce per-sample estimates of: (i) decay rates $\hat{\lambda}^k$ for each of the $K{=}16$ fluorophore components, with output biases initialised to rates log-spaced between the slowest and fastest observable rates (the inverse window and frame durations), (ii) fluorophore abundances $\hat{w}^k$, and (iii) peak amplitudes $\hat{\alpha}^j$ for each of the $m{=}300$ globally shared Voigt profiles introduced in Section 3.1.

Raman spectra are constructed according to Eq. 1 as a weighted sum of Lorentzian peaks, which is convolved once with a Gaussian of width σ_G representing the instrument response; since convolution is linear, this yields exact Voigt profiles with a shared Gaussian component. The peak centre positions μ^j and per-peak Lorentzian widths γ^j are global parameters shared across all samples in the dataset and optimised jointly with the encoder and decoder. The shared Gaussian width σ_G is likewise a single globally shared scalar representing the instrument response function. All widths are kept above the resolution limit given in Appendix C.1 via a softplus offset. By contrast, the peak amplitudes $\hat{\alpha}^j$ are produced by the decoder on a per-sample basis, allowing the model to express the sample-specific Raman fingerprint within a globally constrained spectral vocabulary.

Fluorophore basis. Each of the $K{=}16$ fluorophore spectra is parameterised as a mixture of $C{=}3$ Gaussians (MoG) in normalised wavenumber space, with learned means, log-scales, and mixture logits. The bases are L2-normalised after evaluation to give unit-energy shapes.

C.3 Implementation Details for RADAR

The RADAR baseline was implemented based on the authors' official source code, utilising the standard model structure for both evaluation settings:

Supervised fine-tuning: In this setting, the standard model served as a pre-trained initialisation. Specifically, the full fine-tuning process was conducted independently for each of the five datasets. In this setup, the model was trained for 400 epochs with a batch size of 64. The optimisation process utilised the Adam optimiser with an initial learning rate of 1×10^{-4}. To ensure efficient convergence, a dynamic learning rate scheduling strategy (ReduceLROnPlateau) was employed, which reduced the learning rate by a factor of 0.5 whenever the validation loss plateaued for 10 consecutive epochs, down to a minimum threshold of 1×10^{-12}. To ensure numerical stability during the calculation of

the Root Mean Square Error (RMSE) loss, a small constant $\epsilon = 10^{-10}$ was added:

$$L = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \epsilon},$$

where $\hat{y}_i$ and y_i denote the i-th spectral bin of the predicted and ground-truth components, respectively, and N represents the total number of spectral bins.

Unsupervised inference: For the unsupervised evaluation, we performed direct inference using the pre-trained weights of the standard model provided by the authors, without any further training or parameter updates.

C.4 Human Cell Dataset: Acquisition and Preprocessing Details

Instrumentation. Raman spectral measurements were performed using a 532 nm continuous-wave laser focused through a 100×/0.9 NA microscope objective onto the sample, with approximately 6.2 mW at the sample (100 mW at the source). Scattered light was collected in backscattering geometry through the same objective and directed through a dichroic mirror; a laser line filter suppressed the excitation wavelength before directing the Raman signal to an Andor Kymera 328i spectrometer, equipped with a 600 grooves/mm diffraction grating and a Newton EMCCD detector. Spectra were acquired in video mode at 100 ms per frame, recording the spectral and temporal dimensions simultaneously. None of these components is specialised beyond a conventional confocal Raman setup, and training LUMOS requires only unlabelled bleaching sequences.

Cell preparation. Human bone marrow mesenchymal stem cells (hMSC-BM) were obtained from Cytion (catalogue no. 300665) and differentiated over 14 days in osteogenic or chondrogenic induction medium to produce bone and cartilage cells, respectively. MCF-7 human breast cancer cells were used as a fourth class without further differentiation.

Processing. Raw acquisition files were loaded from 431 spots across four cell classes. Sequences are cropped to the shutter-open region and dark-subtracted using per-sample offset estimated from the shutter-closed region. We filter samples such that the first frame falls within a specified intensity window, ensuring a consistent starting fluorescence level. In order to ensure there is no contamination between inputs and targets, spots were rejected if they contained fewer laser-on frames than the maximum of 64 input frames used throughout our experiments, plus the 16 frames used to obtain the Raman GT. Spots that failed kinetic quality-control checks (intensity drift, isolated spikes, or flatline behaviour) were additionally excluded, yielding 266 valid sequences. Dark current was estimated per spot from shutter-closed frames preceding and following laser exposure; five spots with no dark frames used the dataset-wide median dark (581.6 counts) as a fallback.

Ground-truth Raman spectra were extracted by averaging the final 16 valid frames per spot on the cleaned sequences, applying IarPLS baseline correction (λ=10^6, 50 iterations), bad-pixel correction (494 pixels corrected across 132 spectra), and Savitzky-Golay smoothing (window 9, order 3). The final dataset comprises 266 sequences across 593 wavenumber channels, split 70/10/20 into train (186), validation (26), and test (54) sets using stratified sampling on cell class.

C.5 Licences of existing assets

The Raman spectra underlying all four simulated datasets were obtained through the `raman_data` library (MIT licence) of RamanBench [40]; each dataset retains its original licence. The Pharmaceutical Ingredients dataset [8] is distributed via figshare (`https://doi.org/10.6084/m9.figshare.27931131`) under CC BY 4.0. The Cancer Cell Metabolites dataset [32] is distributed via Kaggle (`https://www.kaggle.com/datasets/andriitrelin/cells-raman-spectra`) under CC BY-NC-SA 4.0. The Mutant Wheat dataset [34] is distributed via Zenodo (`https://doi.org/10.5281/zenodo.7644521`) under CC BY 4.0. RRUFF spectra [33] are made freely available by the RRUFF Project (`https://rruff.info`), which states no explicit licence; we use them for non-commercial research and cite them as requested by the project. Fluorophore emission spectra were obtained from FPbase [38], whose data are provided free of copyright restrictions with attribution to the original authors. RADAR [27]

was used through the authors' public implementation (`https://github.com/JoelSjoberg/Comprehensive-Raman-Spectral-Analysis-Using-Deep-Learning`), for which no licence is specified.

D Failure Case: Extreme Low-Light Conditions

Figure 6 shows a representative example where LUMOS struggles, corresponding to the limitation discussed in Section 5. The sample is from the Human Cell dataset and was selected by jointly minimising first-frame intensity and maximising spectral angle error relative to the ground-truth Raman spectrum.

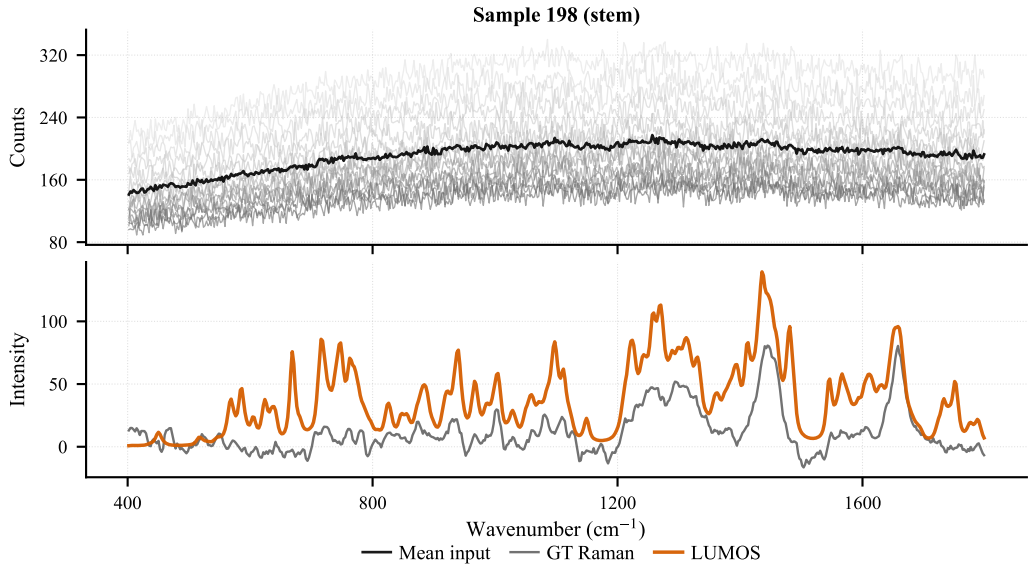


Figure 6: Failure case on the Human Cell dataset under extreme low-light conditions. *Top:* raw input bleaching frames showing intense shot noise. *Bottom:* LUMOS prediction (orange) vs. ground-truth Raman (grey). The model struggles to separate signal from noise in this regime, consistent with the limitations of the linear Gaussian noise approximation when Poisson shot noise dominates.

E Per-Dataset Qualitative and Quantitative Results

Figure legends in this section abbreviate IarPLS as *iairPLS* and refer to the *last* mode of Section 4.2 as *LastFrame* or *Single*.

E.1 Human Cells

Results on the real-world photobleaching dataset comprising bone, cartilage, stem cell, and MCF-7 breast cancer cell classes. Ground-truth Raman spectra are approximated from late-frame averages; residual fluorescence in the target is expected and reflected in the metric distributions.

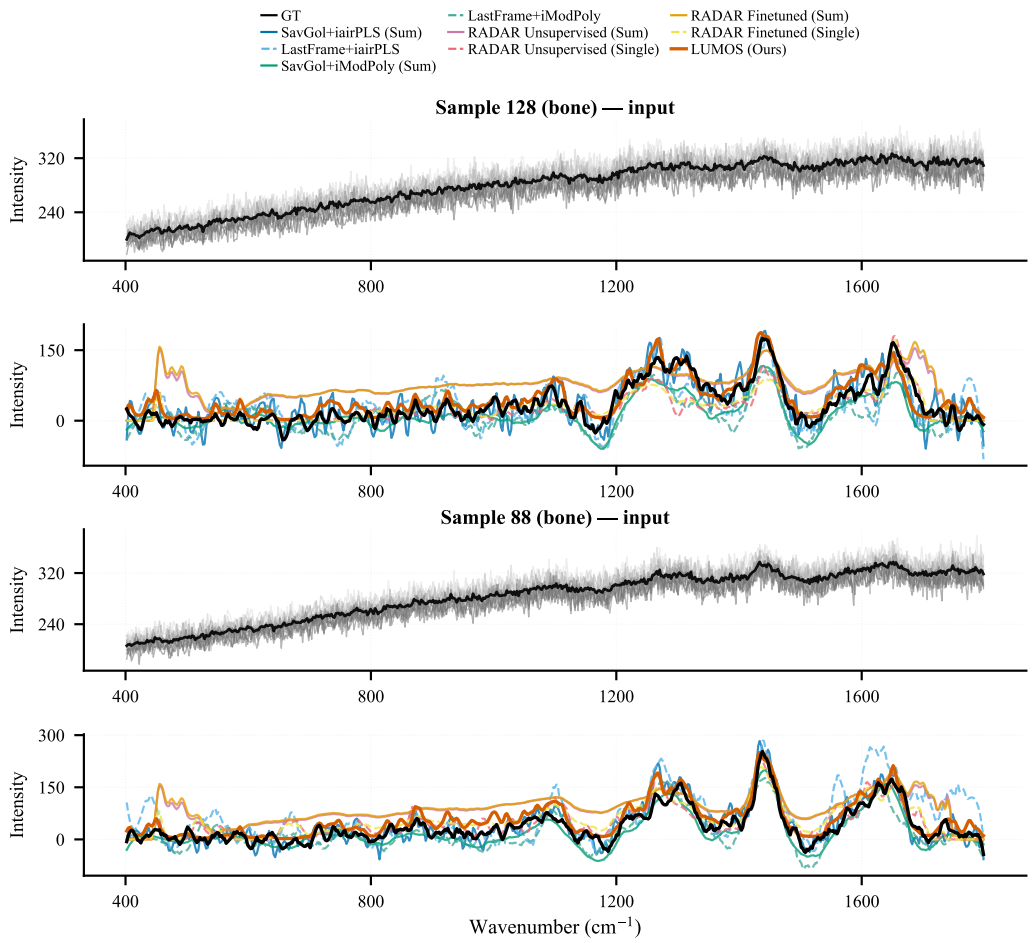


Figure 7: Qualitative comparison of input bleaching frames and extracted Raman spectra for two representative test samples (**Human Cells**). *Odd rows:* raw input time series with mean shown as a solid black line. *Even rows:* predicted Raman spectra for all methods in count space.

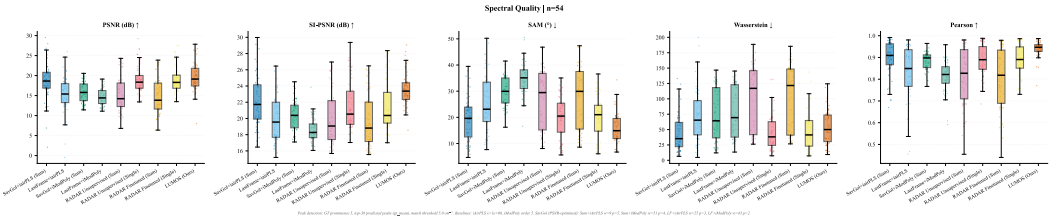


Figure 8: Per-sample distributions of spectral quality metrics across the test set (**Human Cells**). Boxplots show median and interquartile range; jittered points show individual samples.

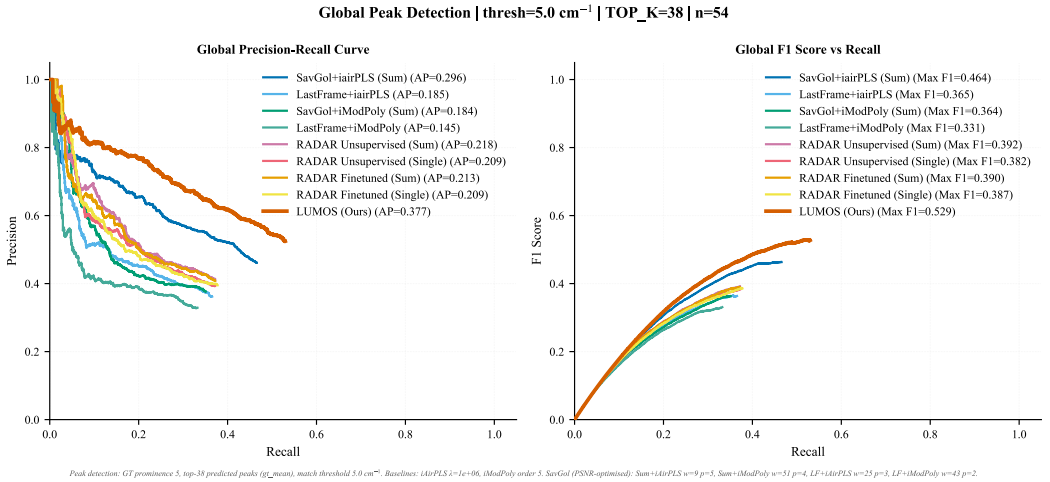


Figure 9: Global precision–recall and F1 curves as a function of detection confidence threshold (**Human Cells**), at a fixed matching distance of $5\,\mathrm{cm}^{-1}$. Area under the PR curve (AUPRC) summarises performance across all confidence levels.

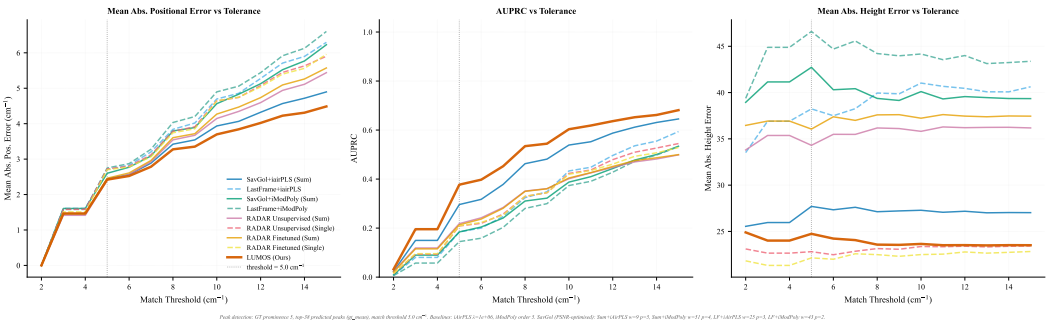


Figure 10: Peak detection performance as a function of matching distance threshold (**Human Cells**). *Left:* mean absolute positional error of true positives. *Centre:* AUPRC. *Right:* mean absolute height error of true positives. The vertical dotted line marks the threshold of $5\,\mathrm{cm}^{-1}$ used in the main table.

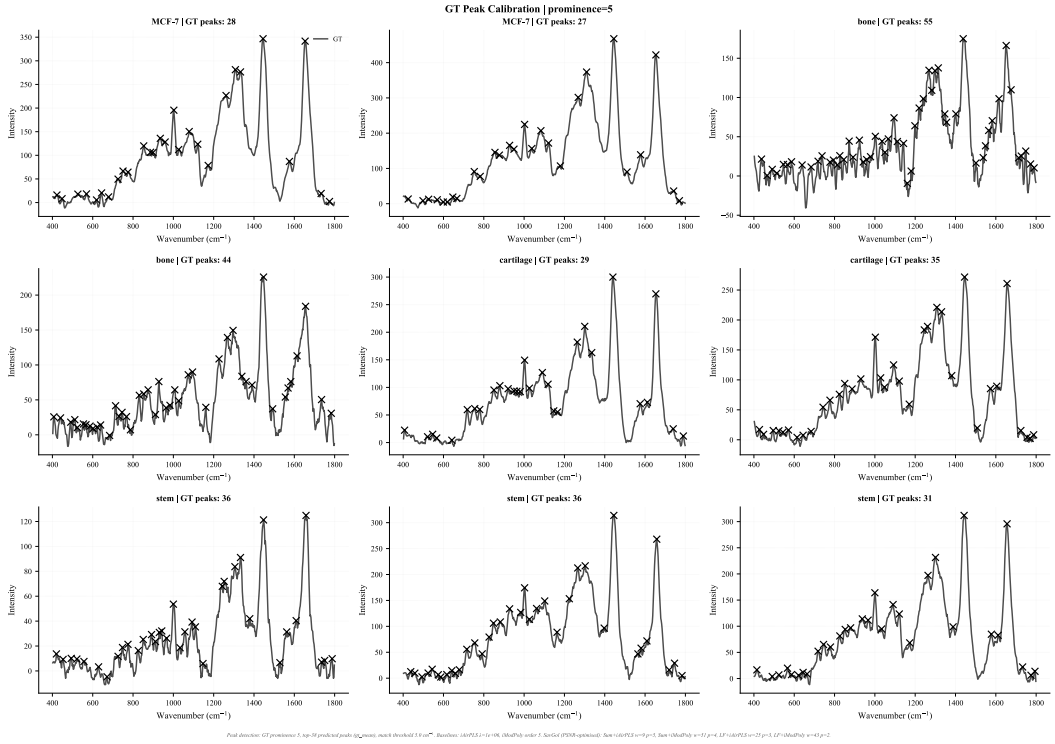


Figure 11: Ground-truth Raman spectra with detected peaks marked for nine representative test samples (**Human Cells**), illustrating the peak prominence threshold used during evaluation.

E.2 Cancer Cell Metabolites

Results on SERS spectra of cancer cell metabolites collected on NH_2-functionalised gold nanourchins across 12 cell line classes.

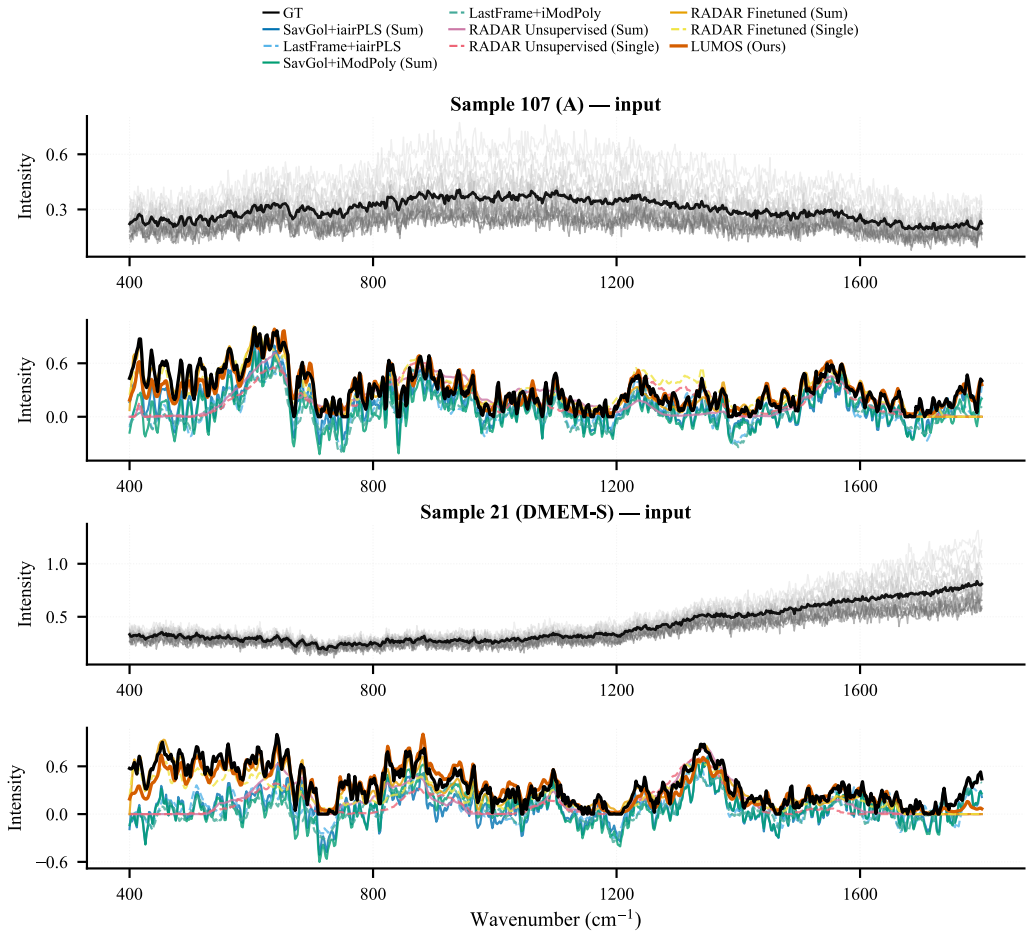


Figure 12: Qualitative comparison of input bleaching frames and extracted Raman spectra for two representative test samples (**Cancer Cell Metabolites**). *Odd rows:* raw input time series with mean shown as a solid black line. *Even rows:* predicted Raman spectra for all methods in count space.

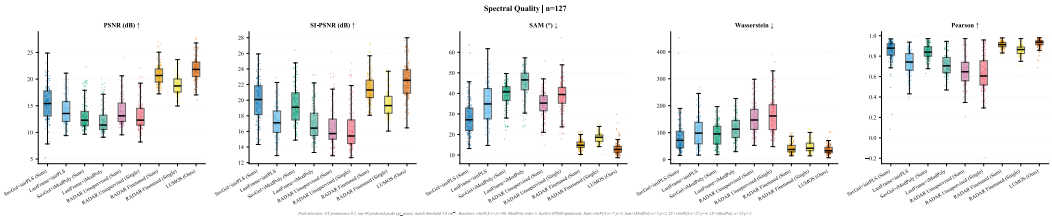


Figure 13: Per-sample distributions of spectral quality metrics across the test set (**Cancer Cell Metabolites**). Boxplots show median and interquartile range; jittered points show individual samples.

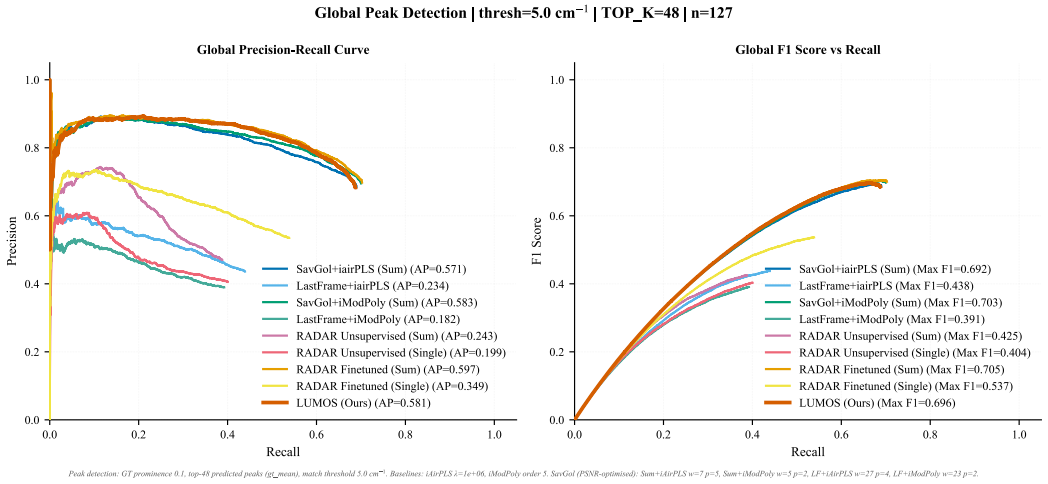


Figure 14: Global precision–recall and F1 curves as a function of detection confidence threshold (**Cancer Cell Metabolites**), at a fixed matching distance of $5\,\mathrm{cm}^{-1}$. Area under the PR curve (AUPRC) summarises performance across all confidence levels.

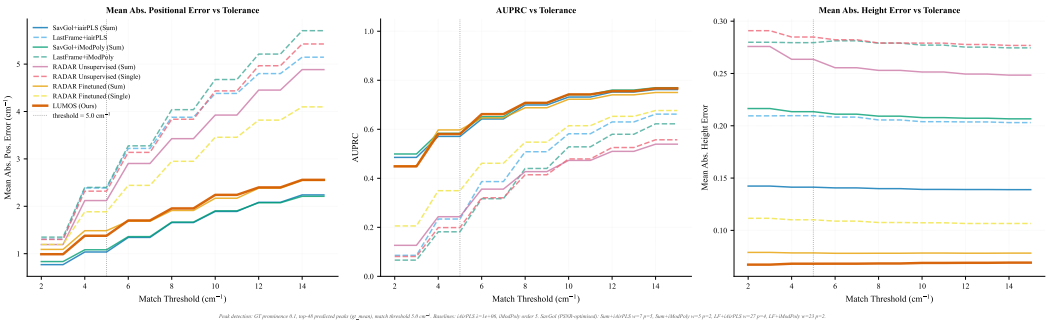


Figure 15: Peak detection performance as a function of matching distance threshold (**Cancer Cell Metabolites**). *Left:* mean absolute positional error of true positives. *Centre:* AUPRC. *Right:* mean absolute height error of true positives. The vertical dotted line marks the threshold of $5\,\mathrm{cm}^{-1}$ used in the main table.

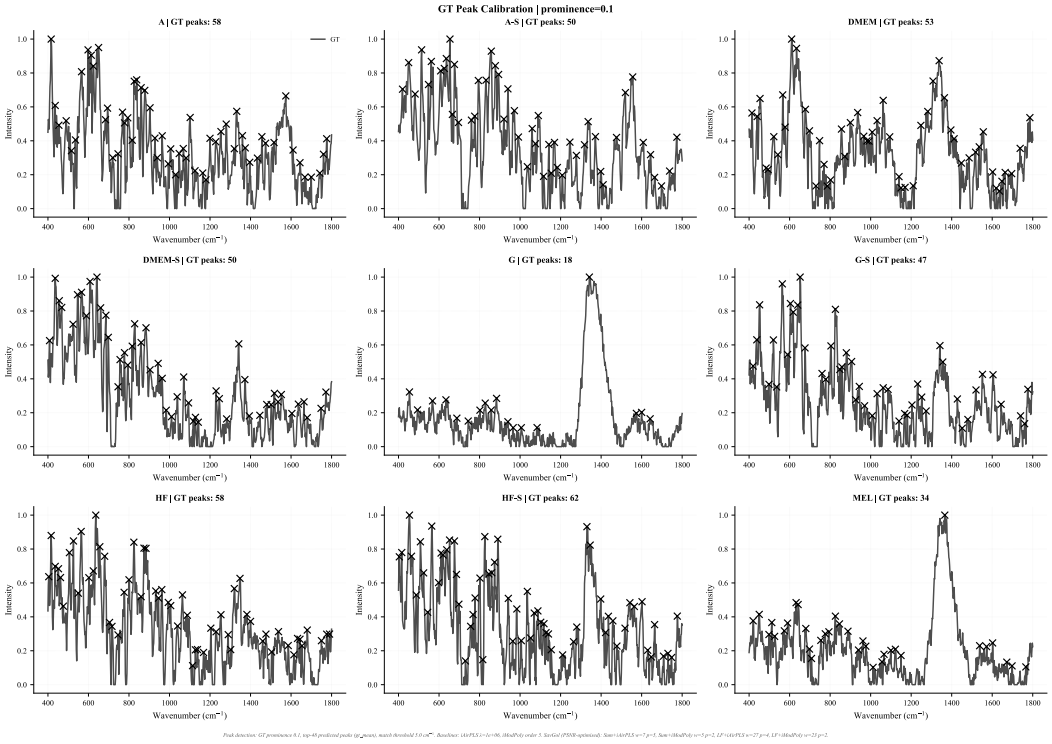


Figure 16: Ground-truth Raman spectra with detected peaks marked for nine representative test samples (**Cancer Cell Metabolites**), illustrating the peak prominence threshold used during evaluation.

E.3 Pharmaceutical Ingredients

Results on the pharmaceutical ingredients dataset comprising 32 classes of organic solvents and reagents used in active pharmaceutical ingredient development.

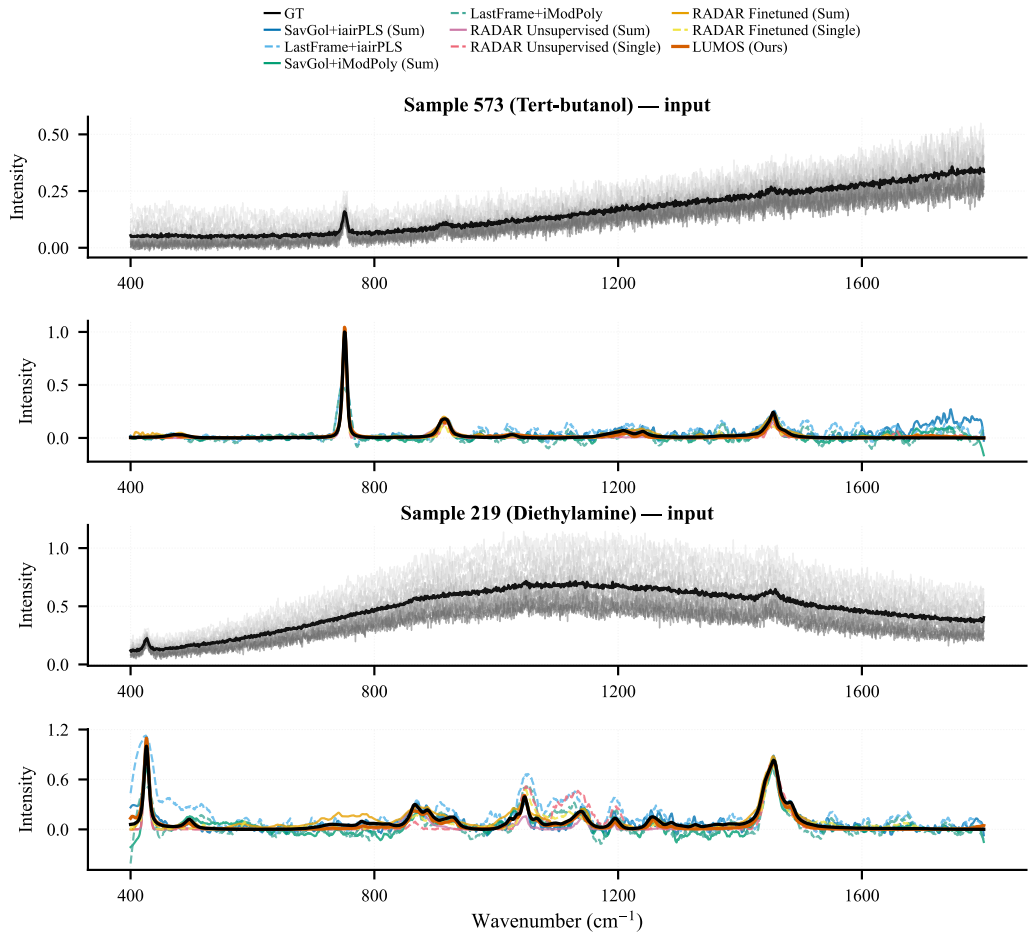


Figure 17: Qualitative comparison of input bleaching frames and extracted Raman spectra for two representative test samples (**Pharmaceutical Ingredients**). *Odd rows:* raw input time series with mean shown as a solid black line. *Even rows:* predicted Raman spectra for all methods in count space.

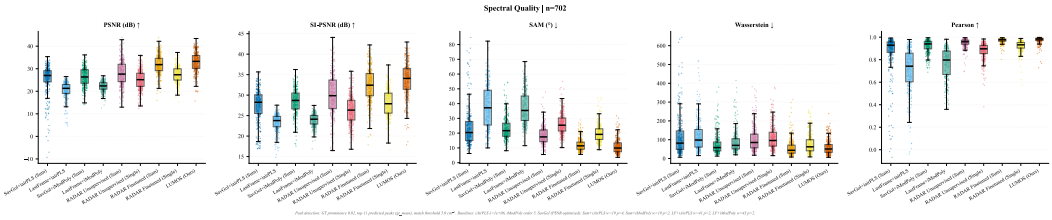


Figure 18: Per-sample distributions of spectral quality metrics across the test set (**Pharmaceutical Ingredients**). Boxplots show median and interquartile range; jittered points show individual samples.

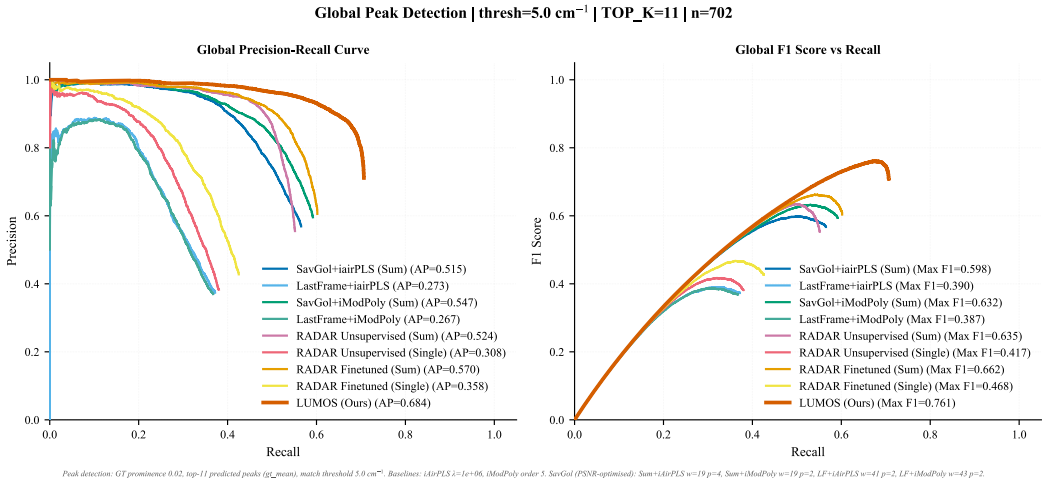


Figure 19: Global precision–recall and F1 curves as a function of detection confidence threshold (**Pharmaceutical Ingredients**), at a fixed matching distance of $5\,\mathrm{cm}^{-1}$. Area under the PR curve (AUPRC) summarises performance across all confidence levels.

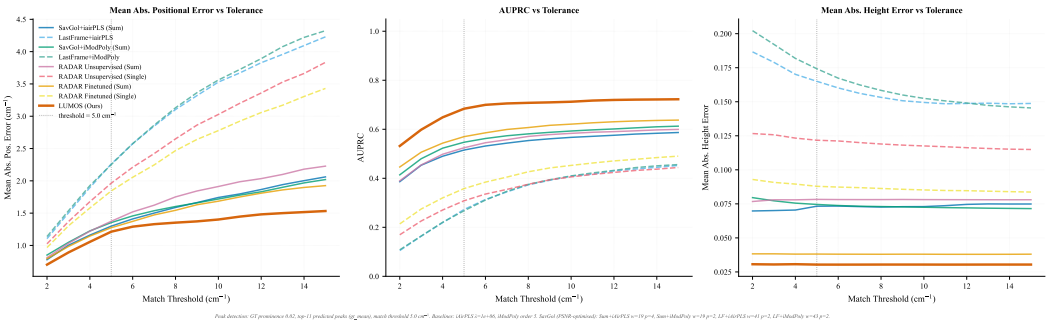


Figure 20: Peak detection performance as a function of matching distance threshold (**Pharmaceutical Ingredients**). *Left:* mean absolute positional error of true positives. *Centre:* AUPRC. *Right:* mean absolute height error of true positives. The vertical dotted line marks the threshold of $5\,\mathrm{cm}^{-1}$ used in the main table.

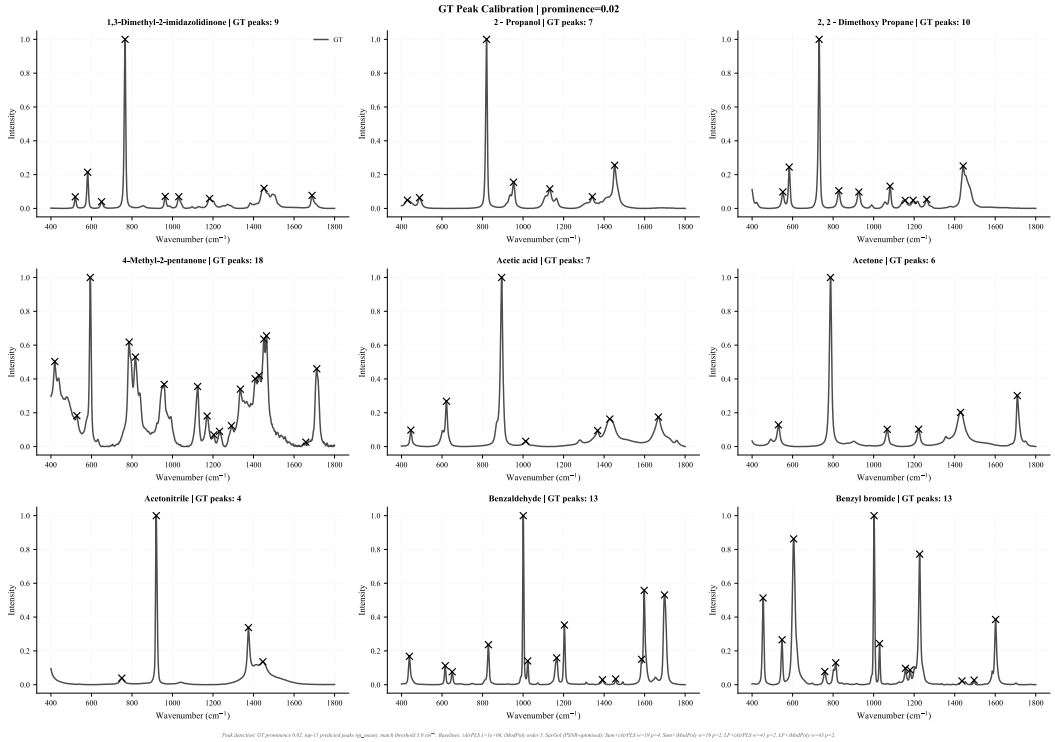


Figure 21: Ground-truth Raman spectra with detected peaks marked for nine representative test samples (**Pharmaceutical Ingredients**), illustrating the peak prominence threshold used during evaluation.

E.4 RRUFF Minerals

Results on the RRUFF mineral database spanning diverse crystallographic structures and chemical compositions.

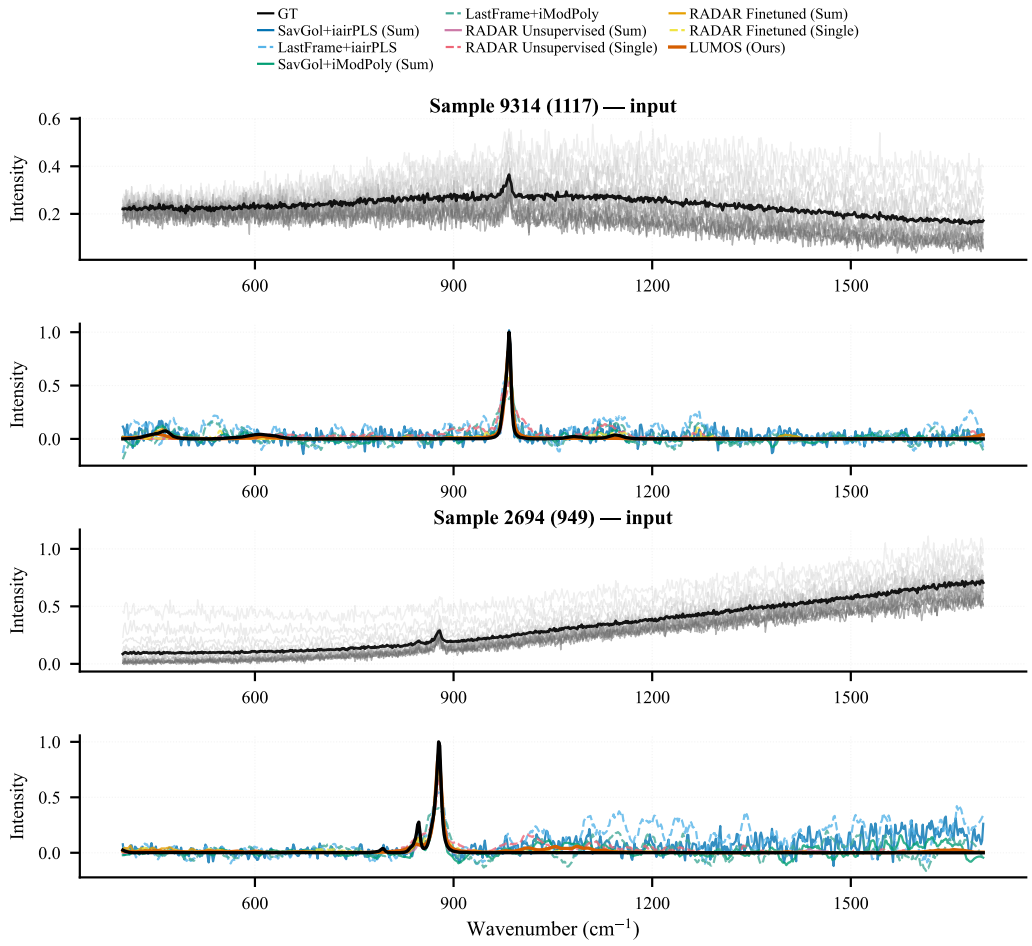


Figure 22: Qualitative comparison of input bleaching frames and extracted Raman spectra for two representative test samples (**RRUFF Minerals**). *Odd rows:* raw input time series with mean shown as a solid black line. *Even rows:* predicted Raman spectra for all methods in count space.

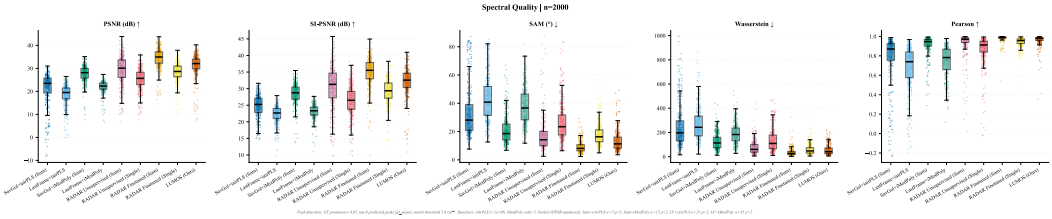


Figure 23: Per-sample distributions of spectral quality metrics across the test set (**RRUFF Minerals**). Boxplots show median and interquartile range; jittered points show individual samples.

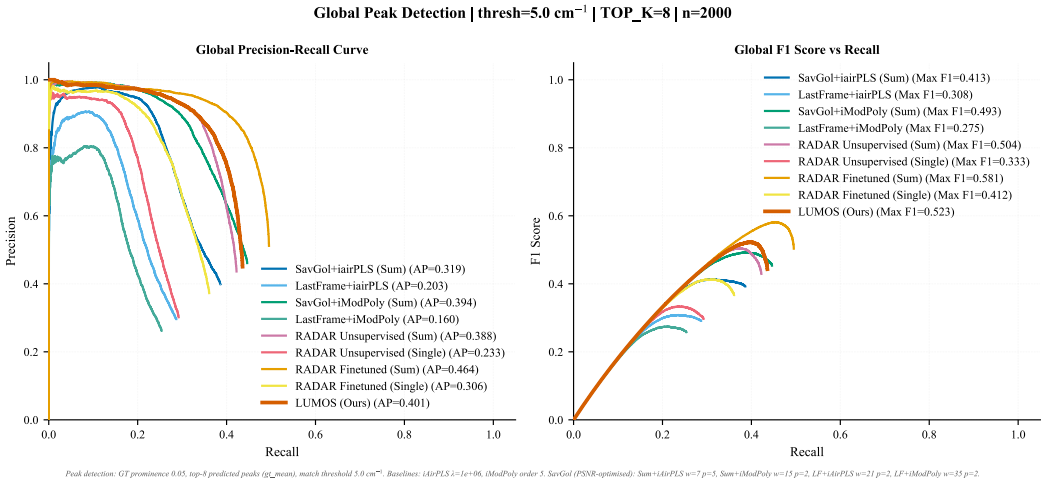


Figure 24: Global precision–recall and F1 curves as a function of detection confidence threshold (**RRUFF Minerals**), at a fixed matching distance of $5\,\mathrm{cm}^{-1}$. Area under the PR curve (AUPRC) summarises performance across all confidence levels.

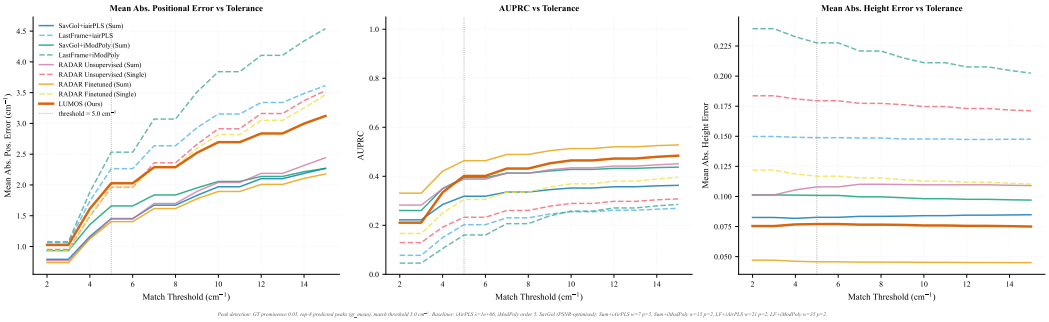


Figure 25: Peak detection performance as a function of matching distance threshold (**RRUFF Minerals**). *Left:* mean absolute positional error of true positives. *Centre:* AUPRC. *Right:* mean absolute height error of true positives. The vertical dotted line marks the threshold of $5\,\mathrm{cm}^{-1}$ used in the main table.

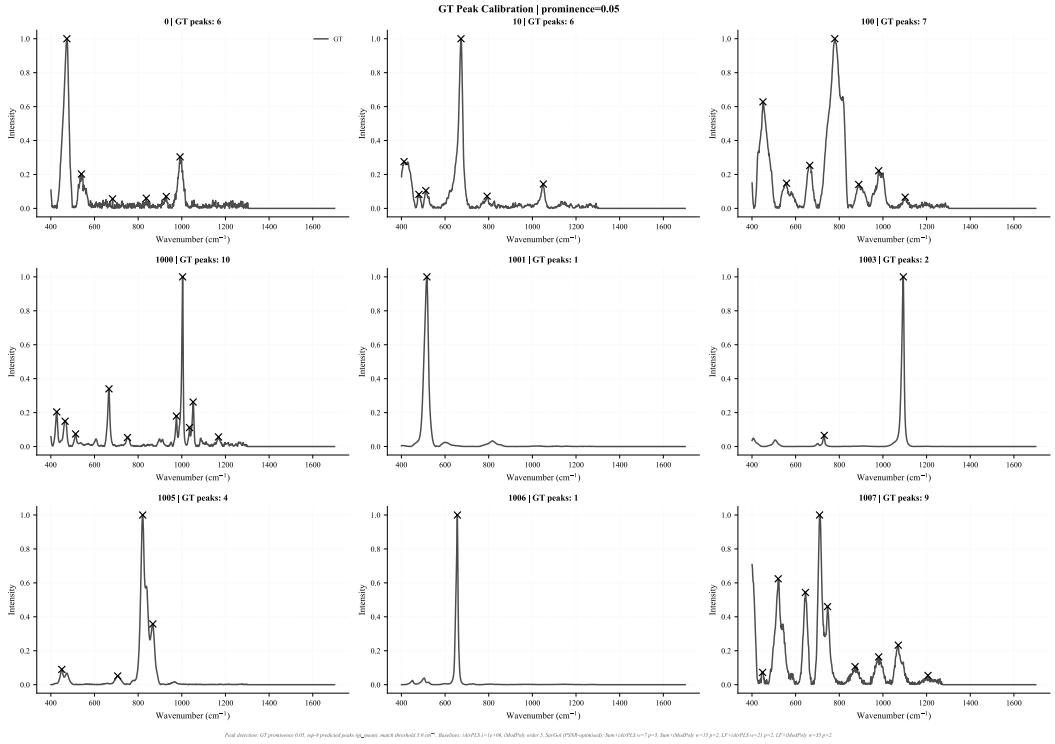


Figure 26: Ground-truth Raman spectra with detected peaks marked for nine representative test samples (**RRUFF Minerals**), illustrating the peak prominence threshold used during evaluation.

E.5 Mutant Wheat

Results on the mutant wheat dataset targeting biochemical shifts in carotenoid and protein-related Raman bands for agricultural phenotyping.

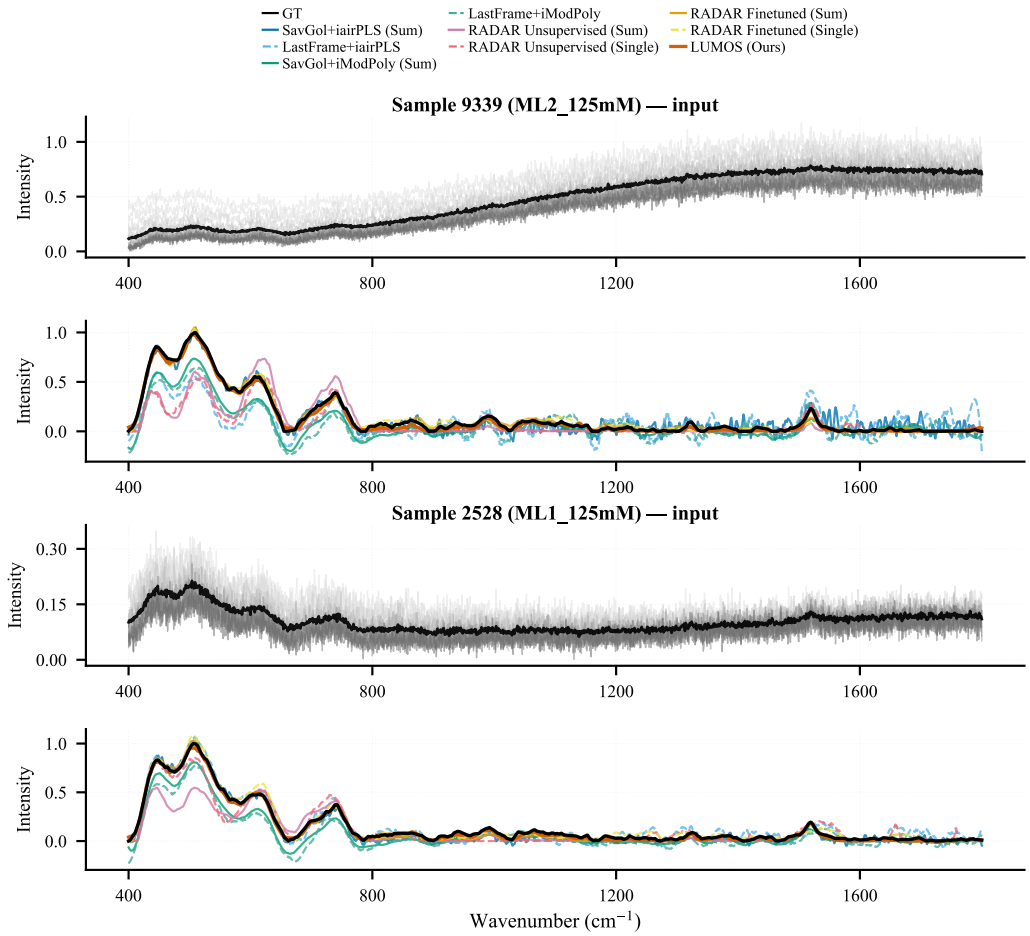


Figure 27: Qualitative comparison of input bleaching frames and extracted Raman spectra for two representative test samples (**Mutant Wheat**). *Odd rows:* raw input time series with mean shown as a solid black line. *Even rows:* predicted Raman spectra for all methods in count space.

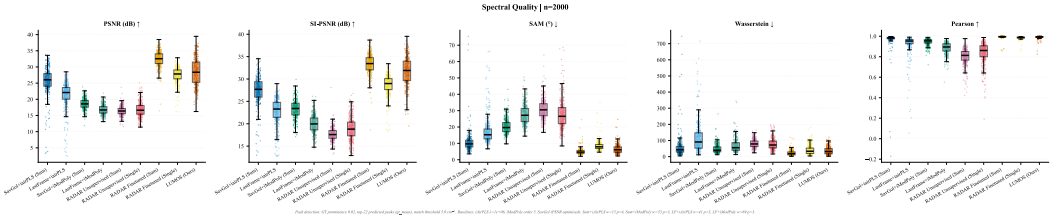


Figure 28: Per-sample distributions of spectral quality metrics across the test set (**Mutant Wheat**). Boxplots show median and interquartile range; jittered points show individual samples.

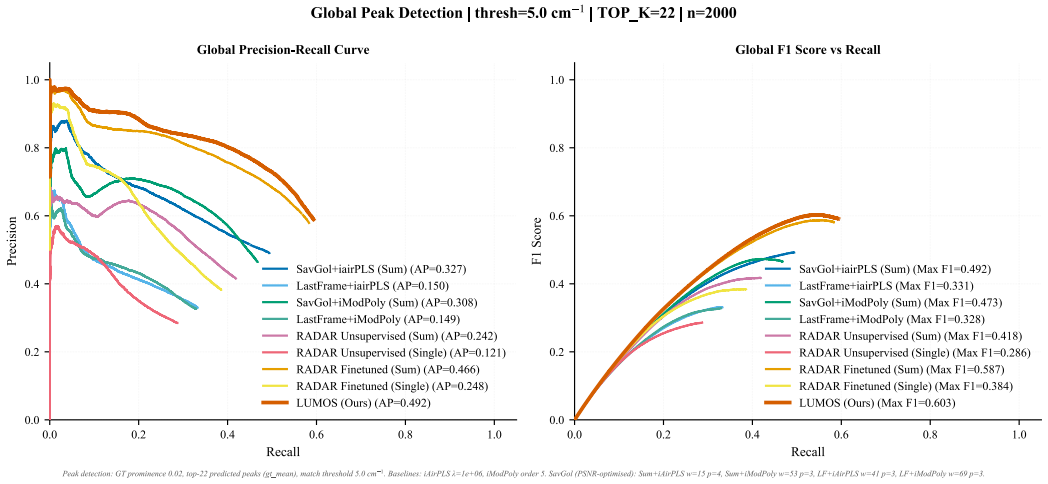


Figure 29: Global precision–recall and F1 curves as a function of detection confidence threshold (**Mutant Wheat**), at a fixed matching distance of $5\,\mathrm{cm}^{-1}$. Area under the PR curve (AUPRC) summarises performance across all confidence levels.

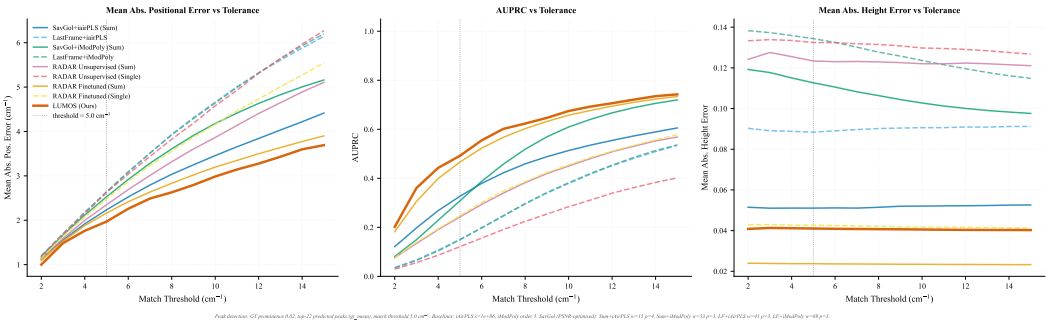


Figure 30: Peak detection performance as a function of matching distance threshold (**Mutant Wheat**). *Left:* mean absolute positional error of true positives. *Centre:* AUPRC. *Right:* mean absolute height error of true positives. The vertical dotted line marks the threshold of $5\,\mathrm{cm}^{-1}$ used in the main table.

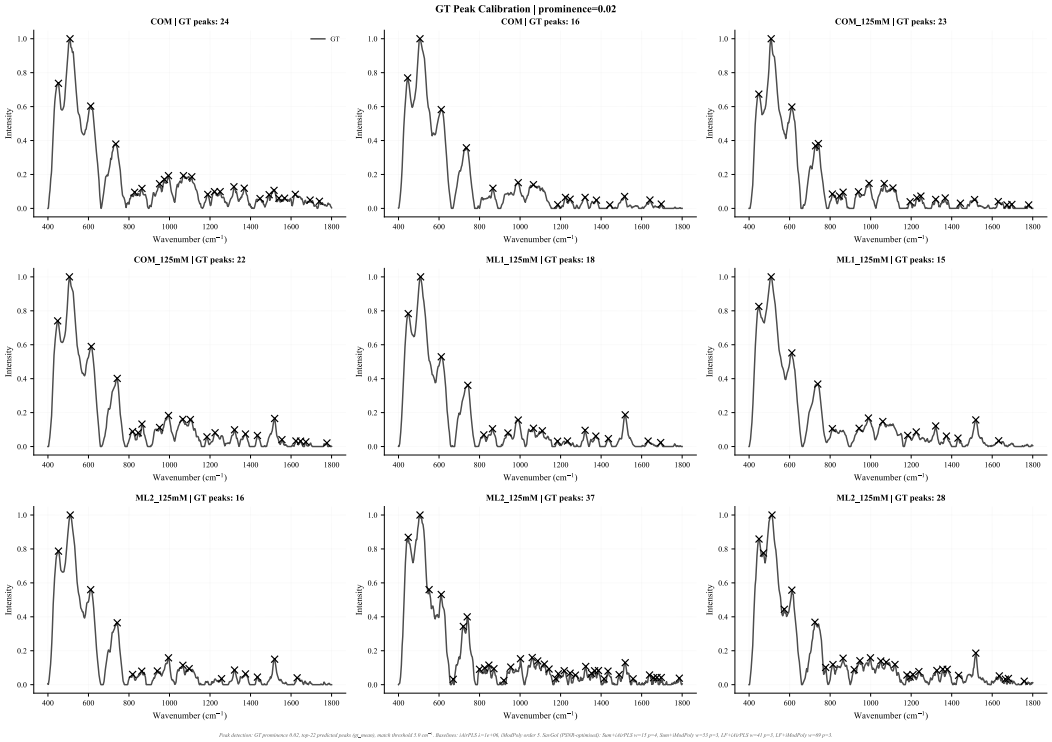


Figure 31: Ground-truth Raman spectra with detected peaks marked for nine representative test samples (**Mutant Wheat**), illustrating the peak prominence threshold used during evaluation.

F Noise Robustness: Extended Results

The following figures provide extended results for the noise robustness ablation described in Section 4. Six Poisson-Gaussian noise conditions of gradually increasing severity were evaluated on the Cancer Cell Metabolites dataset.

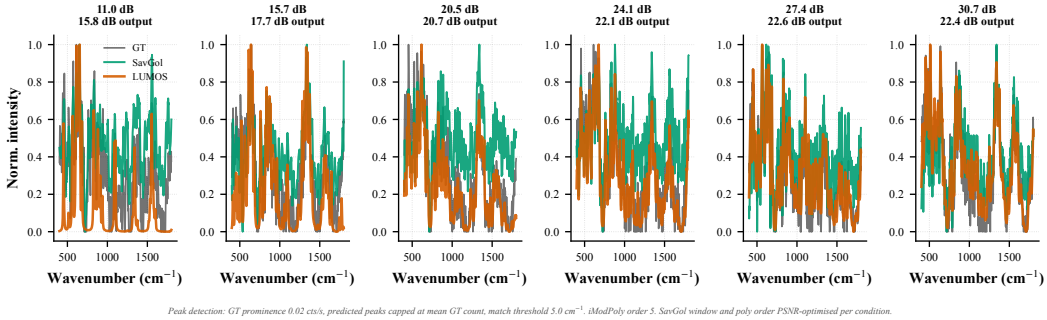


Figure 32: Qualitative comparison of extracted Raman spectra across six noise conditions of increasing severity. Each panel corresponds to one noise level, showing the median-SI-PSNR test sample.

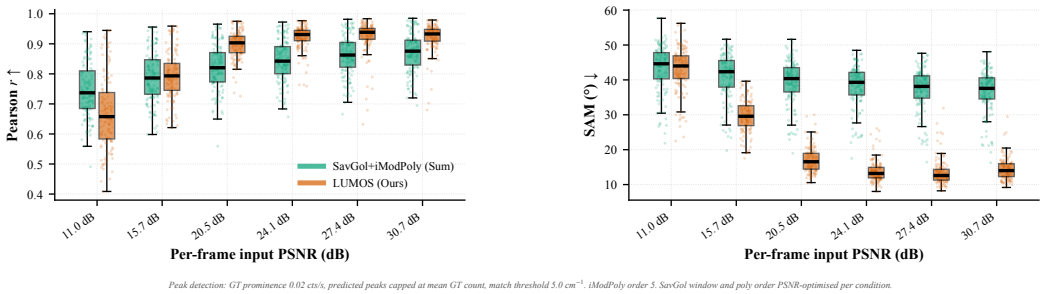


Figure 33: Per-sample Pearson r and SAM distributions across noise conditions. Boxes show the interquartile range; individual points are overlaid.

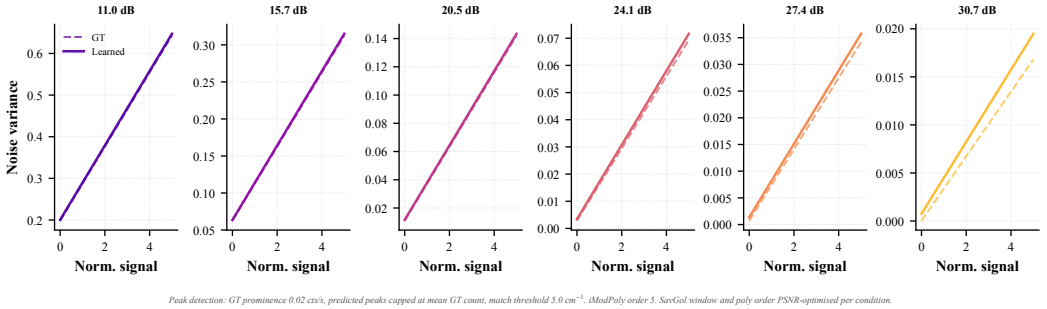
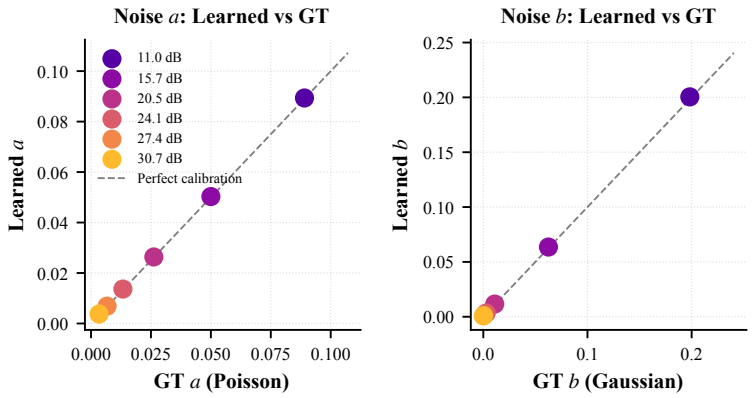


Figure 34: Ground-truth and learned noise variance curves $\hat{\sigma}^2 = \hat{a} \cdot \hat{y} + \hat{b}$ for each noise condition, showing the quality of the estimated noise model parameters (a, b).



Peak detection: GT prominence 0.02 cts/s, predicted peaks capped at mean GT count, match threshold 5.0 cm^{-1}. iModPoly order 5. SavGol window and poly order PSNR-optimised per condition.

Figure 35: Learned vs ground-truth noise model parameters a (Poisson coefficient) and b (Gaussian read-noise floor) across noise conditions. The dashed line indicates perfect calibration.

G Series Length Ablation: Extended Results

The following figures provide extended results for the series length ablation described in Section 4. Input sequence length T was varied across $\{1, 2, 4, 8, 16, 32, 64\}$ frames on the Cancer Cell Metabolites dataset.

Table 3 reports all spectral quality metrics per sequence length for LUMOS and both long-exposure baselines.

Table 3: Effect of sequence length T on the Cancer Cell Metabolites test split (127 spectra). Baselines receive a single long exposure of the same total duration. Values are means over test samples. Best per T in **bold**.

T	Method	PSNR ↑	SI-PSNR ↑	Pearson ↑	SAM ↓
1	LUMOS	**16.56**	**17.62**	**0.76**	**24.36**
	SavGol + iModPoly	12.01	16.34	0.64	46.88
	SavGol + IarPLS	11.45	15.96	0.58	40.47
2	LUMOS	**18.26**	**19.02**	**0.83**	**20.38**
	SavGol + iModPoly	12.30	17.05	0.72	44.88
	SavGol + IarPLS	13.07	16.82	0.68	36.91
4	LUMOS	**19.74**	**20.41**	**0.88**	**17.09**
	SavGol + iModPoly	12.62	17.82	0.77	42.65
	SavGol + IarPLS	14.03	17.68	0.76	34.38
8	LUMOS	**21.00**	**21.84**	**0.91**	**14.56**
	SavGol + iModPoly	12.86	18.61	0.81	40.80
	SavGol + IarPLS	14.89	18.93	0.82	30.41
16	LUMOS	**21.88**	**22.48**	**0.93**	**13.34**
	SavGol + iModPoly	13.11	19.41	0.84	39.16
	SavGol + IarPLS	15.50	20.00	0.85	27.86
32	LUMOS	**22.62**	**23.38**	**0.94**	**12.31**
	SavGol + iModPoly	13.29	20.10	0.86	37.95
	SavGol + IarPLS	16.16	20.79	0.88	25.51
64	LUMOS	**24.12**	**24.83**	**0.95**	**10.42**
	SavGol + iModPoly	13.38	20.57	0.87	37.26
	SavGol + IarPLS	16.76	21.52	0.89	23.41

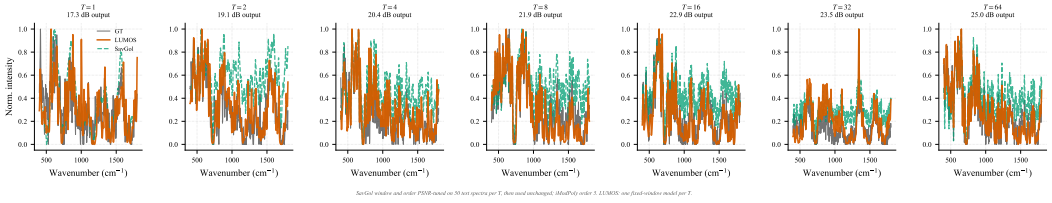
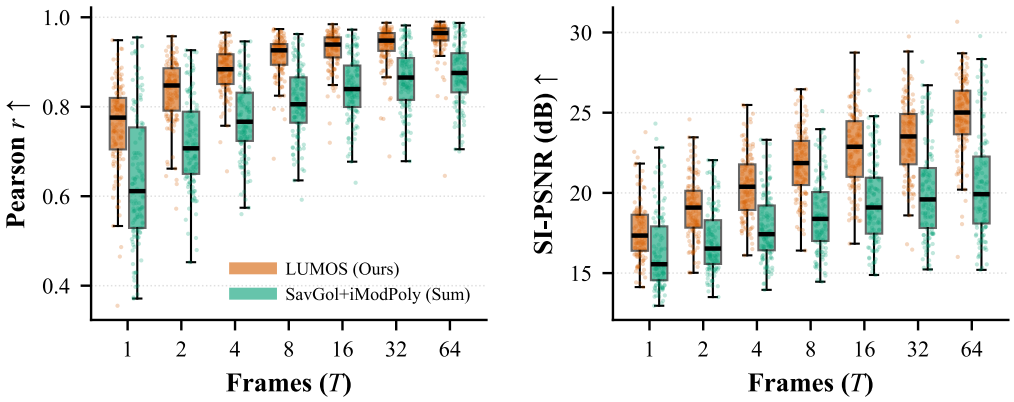


Figure 36: Qualitative comparison of extracted Raman spectra across sequence lengths $T \in \{1, 2, 4, 8, 16, 32, 64\}$. Each panel shows the median-SI-PSNR test sample for that condition.



SavGol window and order PSNR-tuned on 50 test spectra per T, then used unchanged; iModPoly order 5. LUMOS: one fixed-window model per T.

Figure 37: Per-sample Pearson r and SI-PSNR distributions as a function of sequence length T. Boxes show the interquartile range; individual points are overlaid.

H Regularisation Parameter Ablation: Extended Results

The following figures provide extended results for the ℓ_1 regularisation ablation described in Section 4. The weight ρ was varied across three sequence lengths $T \in \{4, 16, 64\}$ on the Cancer Cell Metabolites dataset.

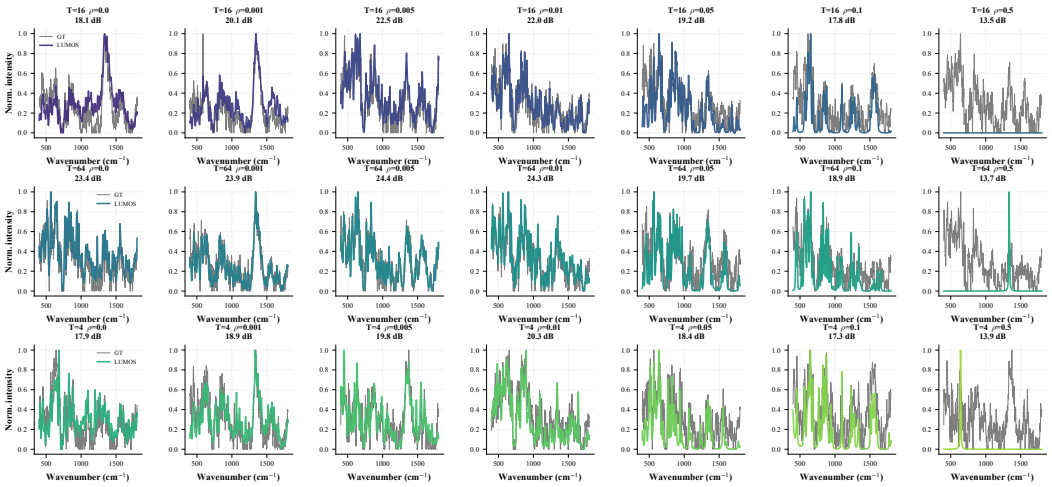


Figure 38: Qualitative comparison of extracted Raman spectra across values of ρ and sequence lengths. Each row corresponds to one value of T; each column to one value of ρ.

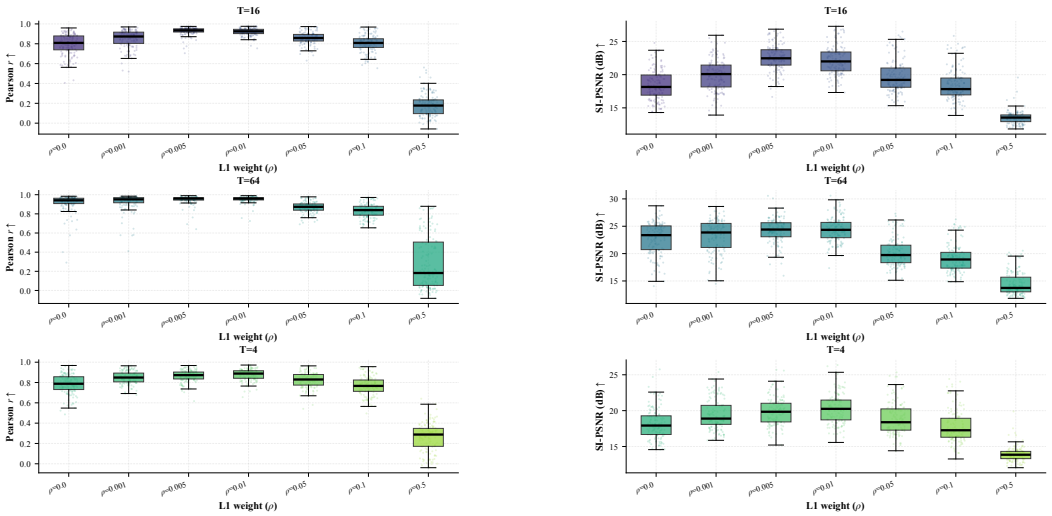


Figure 39: Per-sample Pearson r and SI-PSNR distributions across values of ρ for each sequence length.

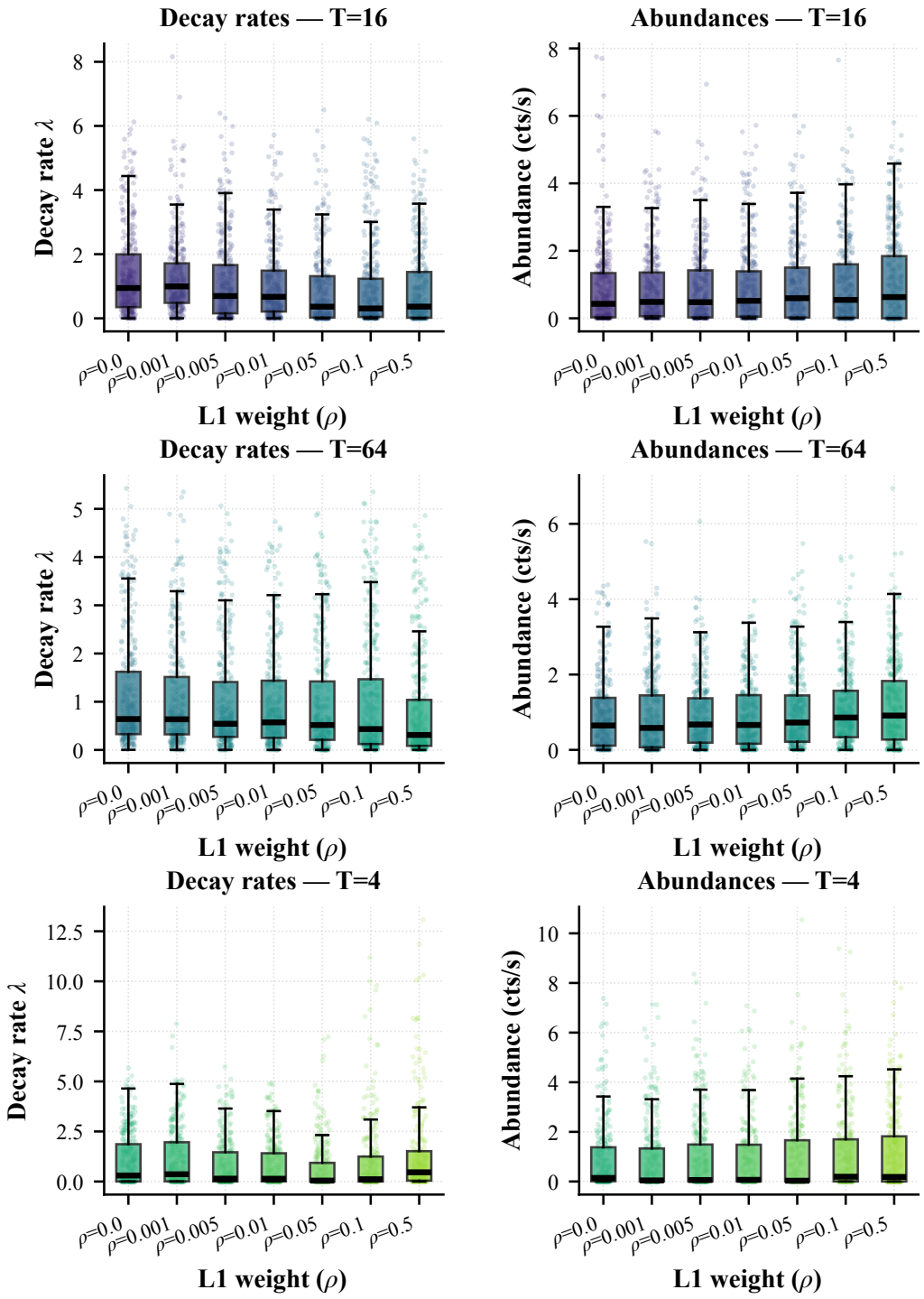


Figure 40: Effect of ρ on the predicted decay rates $\hat{\lambda}^k$ and fluorophore abundances $\hat{w}^k$ across sequence lengths.

I Number of Fluorophore Components

Extending the analysis in Section 4, we retrain LUMOS with different numbers of learnable fluorophores K, keeping all other settings fixed, on two versions of the simulated Cancer Cell Metabolites data: the default 16-fluorophore library (Appendix A) and a smaller 4-fluorophore library, with three fluorophores drawn per sample in both. Small K degrades performance for both libraries, while larger K gives consistently stable results, with no degradation when K exceeds the library size (Table 4, Figures 41 and 42). The default $K{=}16$, which matches the default library and exceeds the number of fluorophores in each sample, lies within this stable range.

Table 4: Effect of the number of learnable fluorophores K for two fluorophore libraries (Cancer Cell Metabolites). Values are means over test samples.

Library	K	Pearson ↑	SI-PSNR ↑	AUPRC ↑
16 fluorophores	4	0.64	17.8	0.54
	8	0.92	22.4	0.57
	16	0.93	22.5	0.58
	32	0.93	22.8	0.58
4 fluorophores	2	0.47	15.8	0.53
	3	0.75	19.8	0.54
	4	0.84	21.4	0.55
	8	0.94	23.2	0.57
	16	0.94	23.1	0.57

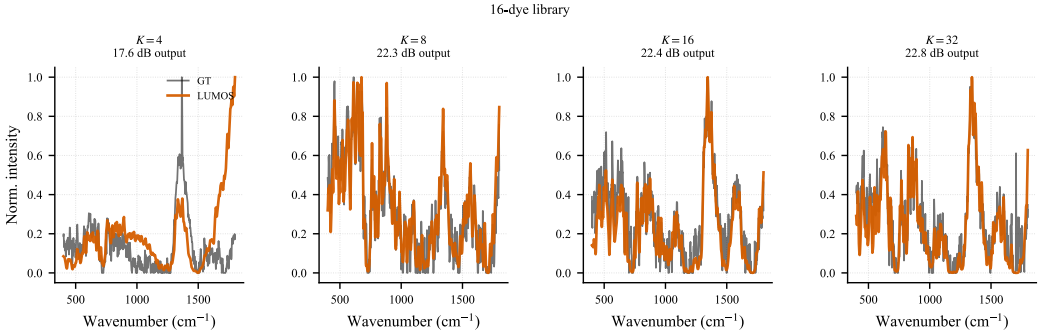


Figure 41: Extracted Raman spectra for different numbers of learnable fluorophores K (16-fluorophore library). Each panel shows the median-SI-PSNR test sample for that model.

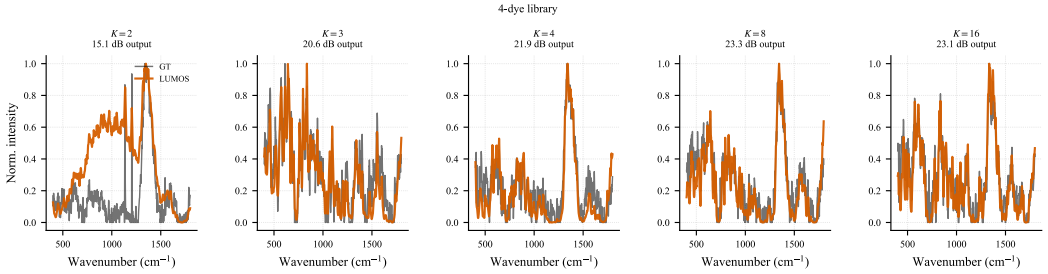


Figure 42: As Figure 41, for the 4-fluorophore library.

J Per-Sample Optimisation Baseline

For each test sample, we fit the forward model of Eqs. 1–5 directly, without encoder or decoder networks. All quantities that LUMOS learns, globally or through its decoder, are optimised independently per sample: Voigt peak positions, Lorentzian widths, the shared Gaussian width and peak amplitudes, the MoG fluorophore bank, decay rates and abundances. The objective is the mean squared error between the reconstruction and the observed series, plus the same ℓ_1 regularisation of the Raman spectrum as LUMOS (ρ=0.01). Parameters are initialised from an untrained LUMOS model of the same configuration. We optimise with Adam (learning rate 3×10^{-2}, 1,500 steps) from three random restarts, keeping for each sample the restart with the lowest objective, so that no ground truth is used for model selection.

In the informed variants, learned quantities are replaced by ground truth: each sample's own fluorophore spectra (true fluorophores), and additionally their decay rates and abundances. Even the latter, which leaves only the Raman spectrum to be fitted, does not outperform LUMOS (Table 5, Figure 43). Notably, all per-sample fits reproduce the observed frames about as well as LUMOS (mean squared reconstruction error 0.026 to 0.029, vs. 0.025), yet, without the true fluorescence parameters, recover a considerably worse Raman spectrum: the frames alone do not determine the decomposition, whereas the prior that LUMOS learns across the dataset does. Fitting with the LUMOS likelihood (Eq. 7, with learnable noise parameters a and b) instead of the mean squared error lowers the uninformed per-sample fits further (Pearson 0.25, and 0.50 with the true fluorophores); only the fully informed fit, which is given all fluorescence parameters, then reaches a Pearson correlation of 0.94, while still detecting peaks worse than LUMOS (AUPRC 0.39 vs. 0.58).

Table 5: Amortised inference (LUMOS) vs. direct per-sample optimisation of the same physics-based forward model (Cancer Cell Metabolites). Rows progressively supply ground-truth fluorescence information to the per-sample fit. Values are means over test samples. Best in **bold**.

Setup	Pearson ↑	SI-PSNR ↑	AUPRC ↑
Per-sample fit	0.43	14.9	0.46
+ true fluorophores	0.69	17.6	0.38
+ true decay rates & abundances	0.92	22.2	0.33
LUMOS (amortised)	**0.93**	**22.5**	**0.58**

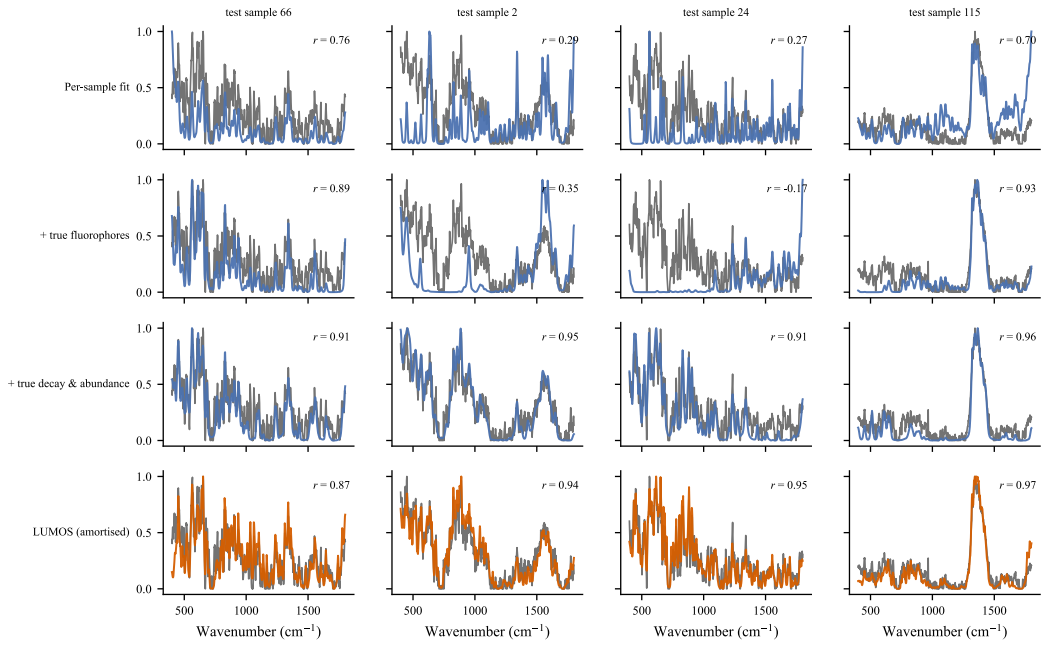


Figure 43: Extracted Raman spectra (colour) against ground truth (grey) for four test samples. Rows: per-sample fits with increasing ground-truth information, and LUMOS. Per-sample Pearson r is shown in each panel.

K Recovery of Fluorophore Parameters

We simulate Cancer Cell Metabolites time series as in Appendix A, using either the full 16-fluorophore library or a library restricted to 4 fluorophores, and train LUMOS with K equal to the library size. Since learnt fluorophores carry no labels, we match them to the ground-truth fluorophores by Hungarian assignment on the Pearson correlation of their spectra, and report the Pearson correlation between matched spectra, abundances and decay rates. As a reference, we replace the learnt bank by the ground-truth spectra (GT dictionary), which removes spectral matching as a source of error. Recovery of abundances and decay rates in the 16-fluorophore setting varies between training runs: an earlier, independent run gave correlations of 0.24 and 0.22 with the learnt bank, and 0.32 and 0.40 with the GT dictionary. In both runs, recovery is poor even with the true spectra, consistent with many near-equivalent decompositions explaining the data equally well when many similar fluorophores are present.

Table 6: Recovery of fluorophore spectra, abundances and decay rates on simulated data (Pearson r after Hungarian matching).

Library	Fluorophore bank	Spectra $r \uparrow$	Abundance $r \uparrow$	Decay rate $r \uparrow$
4 fluorophores	learnt	0.99	0.83	0.74
	GT dictionary	1.00	0.82	0.76
16 fluorophores	learnt	0.86	0.30	0.25
	GT dictionary	1.00	0.38	0.24

L Latent Space and Decay Dynamics: Extended Results

The following figures provide extended analysis of the representations learned by LUMOS on the Human Cell dataset, complementing the UMAP results shown in Figure 5 of the main paper.

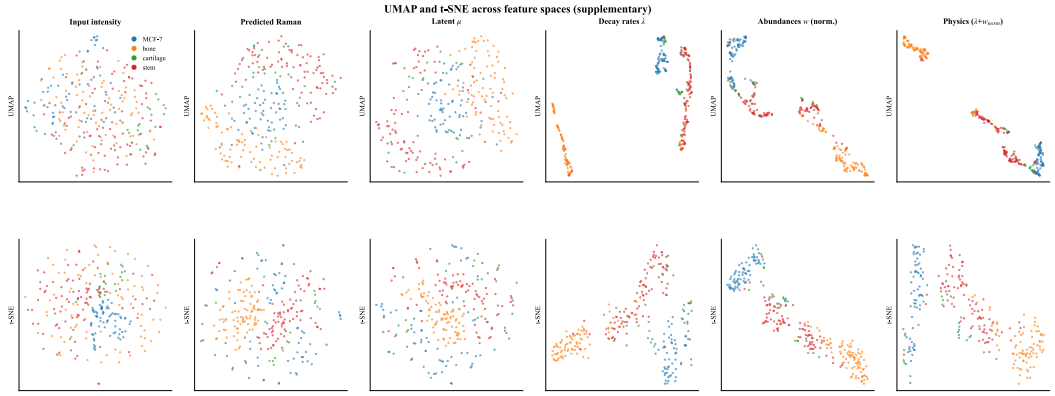


Figure 44: UMAP [37] and t-SNE embeddings of six feature spaces: raw input intensity, predicted Raman spectra, latent mean $\boldsymbol{z}$, decay rates $\hat{\lambda}^k$, normalised abundances $\hat{w}^k$, and the combined physics representation ($\hat{\lambda}^k$+$\hat{w}^k$). Each point corresponds to one test sample, coloured by cell class. Both the decay rates alone and the combined physics representation separate stem cells from the other classes, suggesting that fluorescence dynamics carry discriminative information complementary to the Raman spectrum.

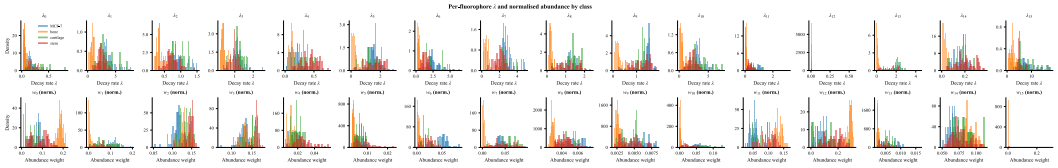


Figure 45: Per-fluorophore distributions of learned decay rates $\hat{\lambda}^k$ (top row) and normalised abundances $\hat{w}^k$ (bottom row) across cell classes. Each column corresponds to one of the K=16 MoG fluorophore components. The separation visible in several components suggests that LUMOS learns class-discriminative fluorescence dynamics without any label supervision.

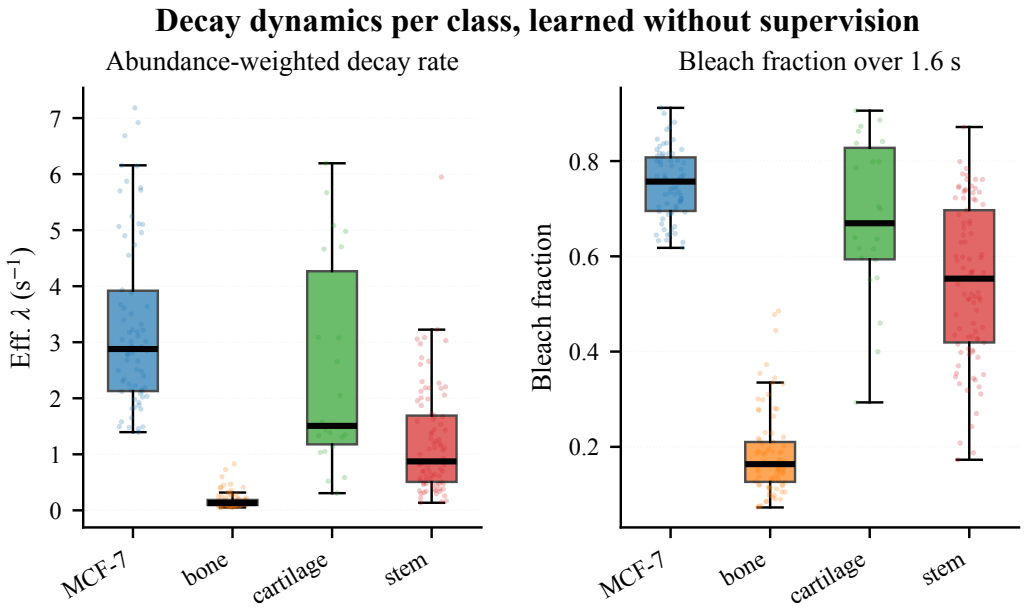


Figure 46: Abundance-weighted effective decay rate (left) and bleach fraction over the acquisition window (right) per cell class, estimated from the decoded physical parameters $\hat{\lambda}^k$ and $\hat{w}^k$. Boxes show the interquartile range; individual points are overlaid. Differences between classes suggest that tissue-specific autofluorescence composition is captured by the model.

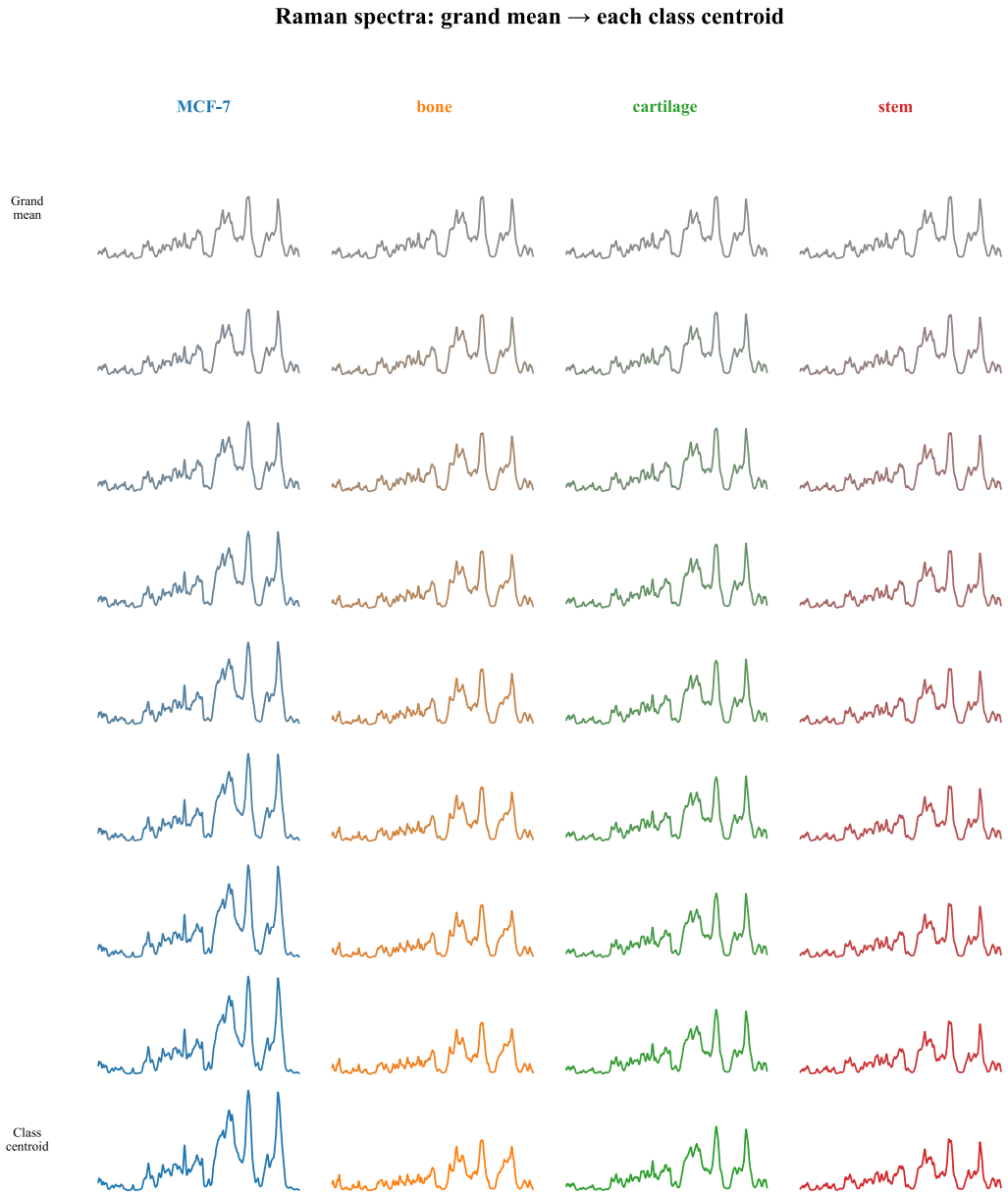


Figure 47: Raman spectra decoded along linear paths from the grand-mean latent code (top row) toward each class centroid (bottom row). Colour saturates from grey at the grand mean to the class colour at the centroid. Differences across columns reveal the class-specific Raman structure learned by LUMOS in a fully unsupervised manner.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the primary contributions: an unsupervised physics-informed VAE for decomposition of incomplete time-resolved Raman spectra, joint modelling of fluorescence decay and noise, and evaluation on both synthetic and real-world datasets. The empirical results presented in Section 4 support these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes an explicit "Limitations and future work" subsection in Section 5 discussing several important limitations of the current approach, including degradation under extreme shot-noise conditions, simplifying assumptions in the Poisson-Gaussian noise approximation, limitations of the Voigt spectral model, and overconfidence in the VAE uncertainty estimates.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents a probabilistic modelling framework but does not make formal theoretical claims requiring theorem statements or mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The probabilistic model, architecture, synthetic data generation process, fluorophore simulation, noise model, and evaluation metrics are described in sufficient detail for understanding and reproducing the experiments. Additional implementation details are provided in the experimental sections and equations in Sections 3 and 4 and in the technical appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The LUMOS implementation and the Human Cell dataset are publicly available at `https://github.com/krulllab/LUMOS`, with scripts and instructions to generate the synthetic datasets, train LUMOS and reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3 and 4 provide detailed descriptions of the model architecture, synthetic data generation process, preprocessing pipeline, evaluation metrics, training settings, hyperparameters, and baseline calibration procedures necessary to reproduce the experiments. Further details can be found in the appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports quantitative evaluation across multiple datasets and metrics. In addition to the aggregate tables in the main paper, Appendix E includes per-sample metric distributions as box plots with overlaid points, and precision–recall and F1 curves across thresholds, showing the variability behind the reported means. The ablation study's Figure 4 directly includes error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.1 specifies the compute resources and training configuration used for the experiments, including GPU type (NVIDIA RTX 3090 with 24 GB memory), training duration, inference time, optimiser settings, and implementation details relevant for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves computational modelling and spectroscopy data analysis and does not violate the NeurIPS Code of Ethics. The work does not involve sensitive personal data or human participants.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents a foundational machine learning method for Raman spectroscopy preprocessing and does not have direct societal deployment implications. Potential misuse risks are limited due to the scientific and low-risk nature of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not release high risk generative models, sensitive datasets, or systems with clear misuse potential requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external datasets, the FPbase fluorophore spectra and the RADAR implementation are cited where used, and their sources, licences and terms of use are listed in Appendix C.5.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces the Human Cell spectral time-series dataset and the LUMOS implementation, both publicly available with documentation and reproduction instructions at `https://github.com/krulllab/LUMOS`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing, human participants, or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work involves no human participants; the only human-derived material is cultured cells, namely hMSC-BM obtained from Cytion (catalogue no. 300665) and MCF-7 cells (Appendix section C.4).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models were not used as part of the proposed methodology, experiments, or scientific contribution of the paper. LLM tools were used only for auxiliary research assistance and editing purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.