Spectral Passport

When Does a Raman Model's Uncertainty Guarantee Survive a Change of Instrument?

Abhay Chakra Sadineni
University of California, Berkeley
abhaychakra@berkeley.edu

November 2026
DOI: 10.5281/zenodo.24210927

Abstract

A Raman calibration model that comes with a conformal prediction interval carries a promise that, on new samples, the true concentration falls inside the interval at a stated rate. The promise rests on the new samples being exchangeable with the calibration samples. This paper tests what happens to it when the spectrometer changes. It uses a public collection of 2,261 Raman spectra of glucose, acetate and magnesium sulfate mixtures measured on eight commercial instruments, 45 of the mixtures on all eight. For each instrument it trains a partial least squares model, calibrates a split conformal interval at 90 percent nominal coverage using one spectrum per mixture, and applies the result on each of the other seven. Across the 168 cross-instrument routes (56 instrument pairs, three analytes), median empirical coverage falls to 0.47, half the routes cover fewer than half of the test spectra, and fewer than one in five reach 0.85. A simple classifier separates every pair of instruments perfectly from a single preprocessed spectrum, so the change is never subtle. Because the instruments' spectra do not overlap at all, weighted conformal prediction, which needs overlap, returns intervals of infinite width for every test sample. Recalibrating with ten labelled spectra from ten distinct mixtures on the target instrument restores coverage on most routes (median 0.93, minimum 0.68) at a cost, since the median interval becomes 1.9 times wider than the full concentration range, and only one route in seven yields an interval narrower than that range. The paper proposes a decision rule, accept, recalibrate or abstain, and a portable record, the Spectral Passport, that binds a prediction to the instrument, the calibration evidence and the decision taken.

Keywords: Raman spectroscopy; conformal prediction; calibration transfer; distribution shift; process analytical technology; uncertainty quantification.

1 Introduction

Process analytical technology depends on a quiet assumption. A model built on one spectrometer will be used on another, whether a replacement probe, a second production line or a contract manufacturer's instrument. Chemometrics has studied this problem for decades under the name calibration transfer, and it has well-tested remedies such as direct and piecewise direct standardisation [1, 2]. Machine learning has since added a second layer. Models now arrive with uncertainty estimates, and conformal prediction offers intervals with a finite-sample coverage guarantee that needs no distributional assumption beyond exchangeability [3–5].

The two layers meet at a point nobody seems to check. When the instrument changes, is the guarantee still true? If it is not, the interval does worse than nothing. It tells an operator, or an automated release decision, that the uncertainty is known when it is not.

This paper asks the question on public data and gets a clear answer. The guarantee fails on most routes as soon as the instrument changes. The tools designed for distribution shift do not rescue it, for a reason that is structural rather than statistical. A small amount of target data does restore it, but the restored intervals are too wide to act on. The useful output is therefore not a better interval. It is a decision about whether to trust any interval at all, recorded in a form that travels with the measurement.

2 Data

The data are the eight-spectrometer bioprocess analyte collection released by Lange et al. [6] and included in RamanBench [7]. Mixtures of glucose, sodium acetate and magnesium sulfate at concentrations typical of *E. coli* fermentation were prepared and measured on eight instruments, with reference values from an enzymatic analyser. The instruments are two Anton Paar systems at 532 and 785 nm, a Kaiser RXN1, a Metrohm, a Mettler Toledo, a Tec5, a Timegate and a Tornado. Each instrument's data come with a published split into training, validation and test sets. In total there are 2,261 spectra, from 133 (Timegate) to 399 (Metrohm) per instrument, of 158 distinct mixtures. Between 53 and 131 mixtures were measured on each instrument, and 45 on all eight. On six of the eight instruments each mixture was measured about five times. The Anton Paar 532 nm system, for example, has 270 spectra of 53 mixtures. The published splits keep the replicates of a mixture together, so each validation set holds between 11 and 27 mixtures and each test set between 10 and 26.

The instruments do not share a wavenumber axis. Every spectrum was linearly interpolated onto a common grid from 400 to 1,800 cm^{-1} at 2 cm^{-1} (701 points), the range covered by all eight. Each spectrum was then converted to its Savitzky-Golay first derivative, using a 15-point window and a second-order polynomial [8], and standardised to zero mean and unit variance across its own points, a standard normal variate step [9].

3 Method

3.1 Source models

For each instrument and analyte, a partial least squares model [10] was trained on the instrument's training set. The number of latent variables, from 1 to 15, was chosen by five-fold cross-validation within the training set, with folds grouped by mixture so that replicates of a mixture never sit on both sides of a fold. The validation set plays no part in fitting or selecting the model. The chosen models used between 1 and 13 components.

3.2 Split conformal intervals

The conformity score is the absolute residual. Replicate spectra of one mixture are not exchangeable with spectra of another, so calibration uses one spectrum per mixture, drawn at random from the source instrument's validation set. With n such residuals, the interval for a new prediction $\hat{y}$ is $\hat{y} \pm \hat{q}$, where $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest residual and $\alpha = 0.10$. If mixtures are exchangeable and replicates of a mixture are exchangeable with one another, the interval covers a new spectrum's true value with probability at least 0.90 [4]. With n between 11 and 27, the guarantee is conservative, because the required order statistic is close to the largest residual. The choice of replicate was repeated 20 times and coverage averaged.

3.3 Routes

A route is a source instrument, a target instrument and an analyte. The source model and its calibrated interval are applied unchanged to the target instrument's test set, and empirical coverage is the fraction of target test samples whose reference value falls inside the interval. There are 168 cross-instrument routes and 24 same-instrument cases, which serve as controls.

3.4 Weighted conformal prediction

When test inputs are drawn from a different distribution than calibration inputs but the relationship between input and output is unchanged, reweighting calibration scores by the likelihood ratio restores coverage [11]. The ratio was estimated with a logistic regression on the first 20 principal components of the pooled source and target spectra, trained to separate source spectra from unlabelled target training spectra. The same classifier's five-fold cross-validated area under the curve serves as a shift diagnostic.

3.5 Recalibration

A recalibrated interval keeps the source model fixed and replaces the calibration set with k labelled spectra from the target instrument's training set, one from each of k mixtures chosen at random, for $k = 10$ and $k = 20$. Each setting was repeated 20 times and coverage averaged.

4 Results

The guarantee fails on most routes. Figure 1 shows coverage by instrument pair, averaged over the three analytes. Across the 168 cross-instrument routes, median coverage is 0.47 and mean coverage 0.49. Only 19 percent of routes reach 0.85. Half fall below 0.50. The model's root mean squared error grows by a median factor of 2.7 (interquartile range 1.7 to 5.0) relative to its error on the source instrument's own validation set.

The shift is never hidden. The domain classifier separates every one of the 56 instrument pairs with a cross-validated area under the curve of 1.00. A single preprocessed spectrum identifies the instrument that produced it. Any deployment can therefore detect an instrument change from unlabelled data alone, before a single wrong interval is issued.

Weighted conformal prediction does not help, and should not be expected to. Because the classes are perfectly separable, the estimated likelihood ratios on target test points are so large that the weighted quantile falls on the point mass at infinity. Every one of the target test samples on every route receives an infinite interval. Coverage is formally 1.00 and informationally zero. The deeper problem is support. Weighted conformal prediction reweights the calibration set toward the kinds of input the deployment will see, which works only if the source instrument could have produced such inputs. When no source spectrum resembles any target spectrum, there is nothing to reweight, and no reweighting of the source calibration set can repair the interval. Whether the relationship between spectrum and concentration also differs between instruments is a separate question, which these tests cannot settle.

Ten target spectra restore validity on most routes but not usefulness. With $k = 10$ labelled target spectra from ten mixtures, median coverage across the 168 routes is 0.93, the minimum is 0.68, and 82 percent of routes reach 0.85, while a third remain below 0.90. Twenty spectra change little (median 0.93, minimum 0.68). But the intervals are wide. The median recalibrated interval is 1.9 times the full range of the analyte in the target test set, and only 23 of the 168 routes, 14 percent, produce an interval narrower than that range. Figure 2 shows the trade. Recalibration

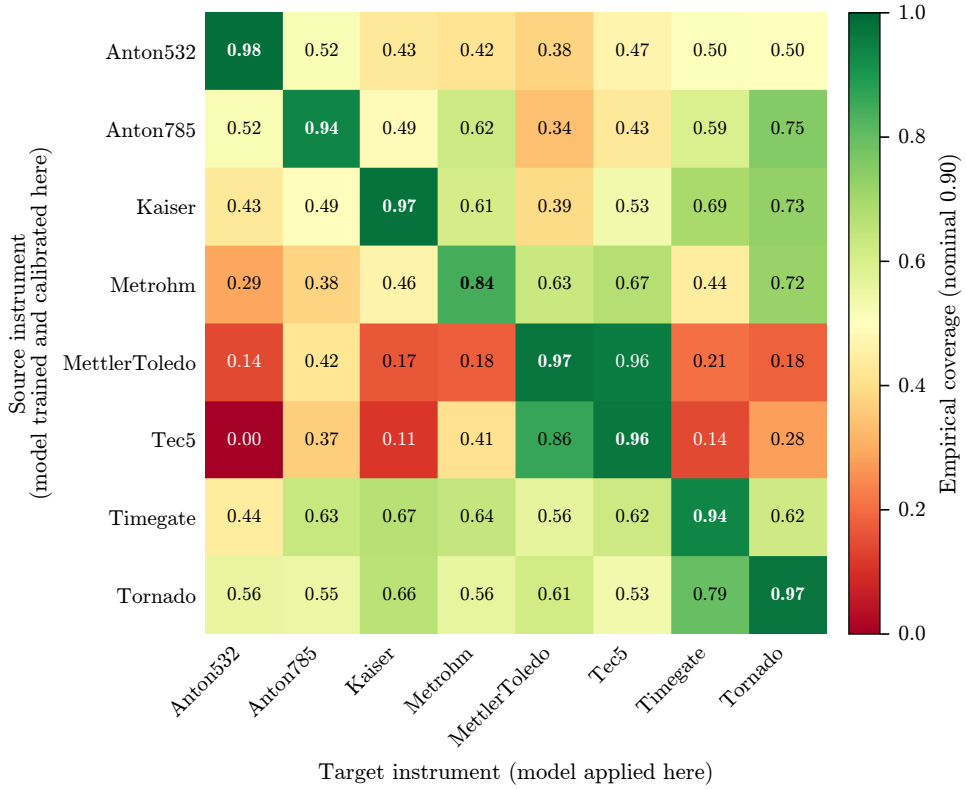


Figure 1. Empirical coverage of a 90 percent split conformal interval when a model trained and calibrated on one instrument (rows) is applied on another (columns). Averages over glucose, acetate and magnesium sulfate. The diagonal is the same-instrument control.

moves most routes to near-nominal coverage and to the right, past the point where the interval says anything.

Even the same instrument is not always exchangeable. The same-instrument controls reach a median coverage of 0.97, above nominal because the small calibration sets make the interval conservative, but they range from 0.71 to 1.00, and 17 percent fall below 0.85. With 10 to 26 test mixtures per instrument, coverage moves in steps of roughly 0.04 to 0.1, which accounts for part of the spread. The cause is in the published splits. For the Metrohm instrument, mean glucose concentration is 1.31 in the validation set and 3.82 in the test set. A guarantee calibrated on one split and tested on another is already operating outside its assumptions. Anyone benchmarking uncertainty on this collection should know this before reading coverage numbers.

5 Accept, recalibrate or abstain

The results suggest a decision procedure rather than a better estimator.

1. *Detect.* Fit a classifier between the calibration spectra and unlabelled spectra from the deployment instrument. If it cannot separate them (area under the curve below a set threshold, 0.75 here), the source interval may be used as issued, and the decision is to **accept**.
2. *Recalibrate.* If the instruments are separable, require k labelled spectra from the deployment instrument and recompute the interval on them, which is to **recalibrate**.
3. *Test usefulness.* If the recalibrated interval is wider than a stated fraction τ of the operating range, the model is not fit for the decision it is meant to support on this instrument. The

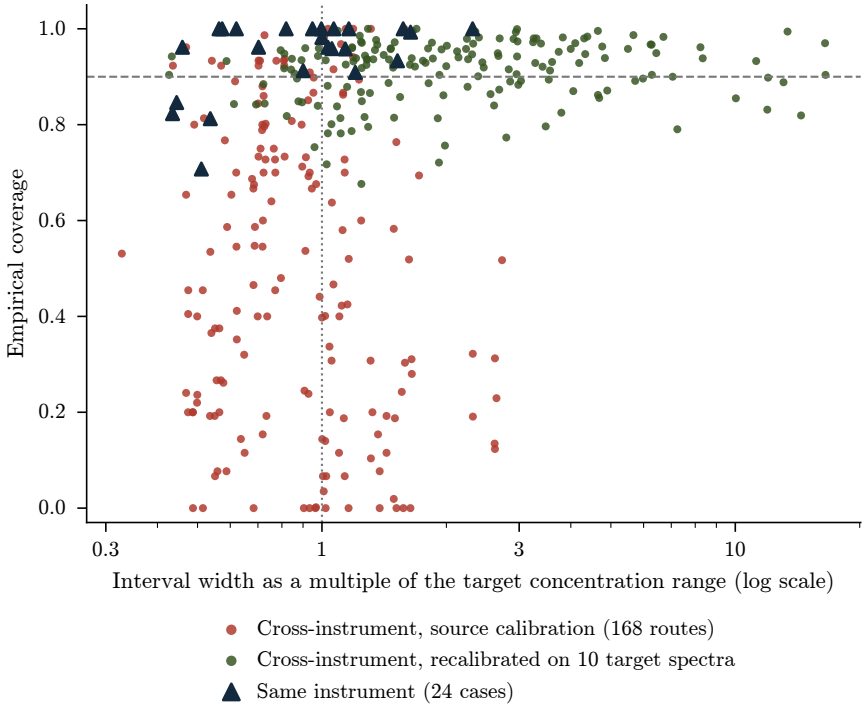


Figure 2. Coverage against interval width for every route. Width is expressed as a multiple of the target analyte's concentration range. At 1.0 the interval is as wide as the full range of values in the test set. The dashed line marks nominal coverage.

decision is to **abstain**, and to retrain or standardise.

Applied to the 168 routes with $\tau = 0.5$, the rule accepts none outright, recalibrates all 168, and then abstains on 166 of them. At $\tau = 1.0$ it abstains on 145. For the partial least squares models used here, the honest answer to whether a model can be moved between these instruments is almost always no. A stronger model, such as the convolutional networks that Lange et al. [6] trained jointly on all eight instruments, would narrow the intervals; the decision rule would then accept more routes, and it would say so on the record.

6 The Spectral Passport

A prediction that has passed through this procedure should carry its evidence with it. The Spectral Passport is a small, signed record attached to each reported value. It names the instrument, the model and its calibration set, the shift diagnostic, the decision taken and the interval actually issued. Figure 3 gives an example. The route and its interval width are taken from the results for the Anton Paar 532 nm glucose model applied on the Tornado. The predicted value is illustrative.

The record does not make a model better. It makes a model's limits visible to whoever relies on the number, be it a quality unit, a regulator or a buyer who never saw the instrument. Chemistry identifies a substance's structure with a standard identifier, the IUPAC InChI [12]. There is no comparable standard for what a measurement of that substance actually showed, on which instrument, with what assurance. The passport is a small step toward one.

```
{
  "schema": "spectral-passport",
  "measurement": {
    "instrument_id": "Tornado", "spectrum_sha256": "...",
    "acquired": "2026-11-13T14:02:00Z" },
  "model": {
    "id": "pls-glucose-anton532", "trained_on": "Anton532", "n_components": 6 },
  "shift_check": {
    "reference_set": "Anton532/val", "classifier_auc": 1.00, "threshold": 0.75 },
  "decision": "recalibrate",
  "calibration": { "set": "Tornado/local-10", "n": 10, "alpha": 0.10 },
  "prediction": {
    "analyte": "glucose", "value": 4.1, "unit": "g/L",
    "interval": [0.0, 19.2], "lower_bound_truncated": true },
  "usefulness": {
    "operating_range": 10.2, "tau": 0.5, "width_ratio": 2.96, "fit_for_use": false },
  "signature": "..."
}
```


Figure 3. An example Spectral Passport. Hashes and the signature are elided.

7 Limits

Only one model class, partial least squares, was tested. Stronger models will narrow the intervals but cannot by themselves create overlap between instruments whose spectra do not overlap. The eight instruments were not measuring identical aliquots in every case, and concentration ranges differ between instruments. The recalibration samples come from the target instrument's training split, which, as Section 4 shows, is not exchangeable with its test split; this biases recalibrated coverage downward on some routes. Coverage is evaluated per spectrum, while the guarantee holds only under exchangeability of mixtures, and with 10 to 27 mixtures per set every coverage estimate is coarse. The shift threshold and τ are set by hand. No standardisation method such as piecewise direct standardisation was tried; combining it with recalibration is the obvious next experiment.

Data and code

The data are public on the Hugging Face Hub under the dataset names beginning `https://huggingface.co/datasets/chlange/RamanSpectraBioprocessAnalystes`, followed by the instrument name (for example, Kaiser). The analysis is a single Python script using NumPy, SciPy, pandas and scikit-learn, with fixed random seeds. The script and the per-route results table are published with this paper on the author's research page.

Funding and interests

No external funding was received. The author declares no conflicts of interest.

References

[1] Wang, Y., Veltkamp, D. J., and Kowalski, B. R. (1991). Multivariate instrument standardization. *Analytical Chemistry*, 63(23), 2750-2756.

[2] Feudale, R. N., Woody, N. A., Tan, H., Myles, A. J., Brown, S. D., and Ferré, J. (2002). Transfer of multivariate calibration models: A review. *Chemometrics and Intelligent Laboratory Systems*, 64(2), 181-192.

[3] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer.

[4] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.

[5] Angelopoulos, A. N., and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494-591.

[6] Lange, C., Borisyak, M., Kögler, M., Born, S., Ziehe, A., Neubauer, P., and Cruz Bournazou, M. N. (2025). Comparing machine learning methods on Raman spectra from eight different spectrometers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 334, 125861. `https://doi.org/10.1016/j.saa.2025.125861`

[7] Koddenbrock, M., Lange, C., Legner, R., Jäger, M., Kögler, M., Cruz Bournazou, M. N., Neubauer, P., Biessmann, F., and Rodner, E. (2026). RamanBench: A large-scale benchmark for machine learning on Raman spectroscopy. arXiv:2605.02003v1, 3 May 2026.

[8] Savitzky, A., and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627-1639.

[9] Barnes, R. J., Dhanoa, M. S., and Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5), 772-777.

[10] Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.

[11] Tibshirani, R. J., Foygel Barber, R., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

[12] Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7, 23.