



Limits and benefits of unsupervised transfer learning for Raman spectroscopy-based bioprocess monitoring

Bianca Hüpf¹ · Alexandra Umprecht^{1,2} · Bence Kozma¹ · Peter Ettinger³ · Oliver Spadiut¹

Received: 1 February 2026 / Accepted: 13 July 2026
© The Author(s) 2026

Abstract

Reliable analyte prediction from Raman spectra is critical for bioprocess monitoring, yet spectral variability across domains, such as different media or scale-up conditions, challenges model generalization. Models for analyte prediction are typically trained solely on labeled source-domain data, often limiting their performance under domain shifts. Transfer learning has been successful in other fields, but is underexplored for spectroscopic data in upstream bioprocesses. To help bridge this gap we assess whether unsupervised transfer learning methods can enhance model generalization by leveraging unlabeled target-domain spectra. We benchmark four unsupervised transfer learning approaches (CORAL, JDOT, TCA, OPP-MMD) against two bioprocess-relevant baselines: a fixed standard preprocessing workflow (REF) and an optimized source calibration based on systematic preprocessing screening (OPP). All methods are evaluated on two bioprocess Raman datasets: Dataset A comprises five glucose-spiked media and is complemented by in-silico spectral perturbations designed to mimic commonly occurring Raman noise types, such as baseline shifts, scattering artifacts, and spectral shifts. Dataset B evaluates scale-up transfer from pilot to production bioreactors for predicting glucose, biomass, and phosphate. Across the investigated cases, a consistent pattern emerged: JDOT and CORAL were most beneficial when across-domain differences dominated, whereas under high within-domain variability transfer learning provided no benefit and even underperformed the baseline. Because these transfer learning approaches can be integrated with standard Raman preprocessing pipelines, unsupervised transfer learning is best viewed as a complementary PAT tool rather than a universal replacement.

Keywords Raman spectroscopy · PAT · Bioprocess modeling · Unsupervised transfer learning · Domain adaptation · Preprocessing · PLS

Introduction

The Process Analytical Technology (PAT) initiative has advanced bioprocess monitoring and control by integrating process sensors, modern computational resources, and chemometric modeling techniques [1–3]. Central to PAT is the real-time, non-invasive measurement of critical process parameters and critical quality attributes using spectroscopic methods such as near-infrared and Raman spectroscopy [4, 5]. Combined with multivariate modeling approaches such as partial least squares regression (PLS), these measurements can support real-time monitoring and control [6–8]. Among chemometric methods, PLS regression is widely used for spectral data modeling due to its ability to handle high-dimensional and collinear data typical for spectroscopy [9, 10]. PLS constructs a latent space by projecting spectral measurements (also termed predictors or features) onto components that maximize their covariance with the

✉ Bianca Hüpf
bianca.mikulasek@tuwien.ac.at

Alexandra Umprecht
alexandra.umprecht@takeda.com

Bence Kozma
bence.kozma@boku.ac.at

Peter Ettinger
peter.ettinger@novartis.com

Oliver Spadiut
oliver.spadiut@tuwien.ac.at

¹ Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorferstraße 1a, 1060 Vienna, Austria

² Baxalta Innovations GmbH, A Member of the Takeda Group of Companies, Industriestraße 67, 1220 Vienna, Austria

³ Novartis Pharmaceutical Manufacturing GmbH, Biochemiestrasse 10, 6250 Kundl, Austria

analyte concentration (also called response or labels), allowing modeling even when the number of spectral variables exceeds the number of samples. A key assumption behind PLS (and many other models) is that the spectra used to build the model and the spectra encountered later in routine use come from the same underlying distribution [11]. In this context, the source domain refers to the setting where the model is developed and calibrated (the available training spectra with known analyte concentrations), while the target domain refers to the new setting where the same model is later applied for prediction (new spectra, typically collected under different conditions). In practice, this assumption is frequently violated: spectra from shake flasks can differ systematically from spectra collected in larger reactors, and additional shifts can arise from changes in medium composition, equipment, or environmental conditions. As a result, models calibrated in one setting can show degraded predictive performance when applied in another [12–14]. Several factors contribute to this mismatch. In practice, spectra reflect more than analyte information: scattering effects, baseline variability, and measurement noise can mask or distort analyte-related signals [15]. If these contributions dominate, PLS may emphasize condition-specific artifacts rather than concentration-related variation, reducing model performance under changing measurement contexts. Spectral preprocessing is therefore routinely applied to suppress artifacts and to highlight analyte-related features. Since many preprocessing techniques and parameter settings exist, selection is still often performed by trial and error [9, 15, 16]. While this workflow is convenient, it can yield models that appear strong in the source domain but generalize poorly to new target spectra. To mitigate such effects, practitioners use sophisticated preprocessing techniques (including automated selection, ensembles, or orthogonal signal correction) [17–19] and established calibration-transfer workflows such as standardization, global models, or slope-and-bias correction [20, 21]. More recently, hybrid workflows combining classical calibration-transfer methods with mechanistic modeling have also been investigated [22]. However, these strategies can be insufficient when the source and target settings differ substantially, resulting in poor model performance even when careful preprocessing is applied [23, 24].

Transfer learning (TL) offers an alternative by incorporating information from both source and target domains to improve predictions under target conditions [25–27]. In the unsupervised setting considered here, only unlabeled target spectra are used during model training, which is an attractive scenario for bioprocesses, where obtaining reference measurements from new setups can be labor-intensive and expensive. TL has been applied successfully in several fields, including geology, agriculture, chemical engineering,

forensics, medical diagnosis, and drug discovery [25, 28–34]. However, despite the prevalence of scale- and condition-induced spectral shifts in bioprocesses, only limited work has explored TL-guided preprocessing selection for Raman data in upstream use cases [35], and systematic comparisons to established TL methods are still missing. Accordingly, we evaluate whether unsupervised transfer learning can improve model performance of PLS-based models under bioprocess-relevant domain shifts, using only unlabeled target spectra. We benchmark four representative methods (CORAL, JDOT, TCA, and OPP-MMD) against two source-only baselines that reflect common PAT workflows: a fixed standard preprocessing chain and an optimized source calibration obtained by systematic preprocessing screening. The evaluation spans two upstream Raman datasets capturing pronounced domain variability in media composition and bioreactor scale. We further stress-test the methods by introducing controlled in-silico perturbations relevant to bioprocess Raman data (white noise, peak shifts, baseline drift) and label-distribution mismatch. Beyond reporting performance, we analyse when unsupervised TL improves upon the source-only baselines, suggesting that gains are analyte- and domain-dependent and can disappear under high within-domain variability. This positions TL as a complementary PAT tool alongside conventional workflows.

Materials and methods

We evaluated all methods on two Raman datasets, restricting analysis to the 400–1800 cm^{-1} region.

Datasets

Dataset A (media transfer)

Dataset A includes spectra of five media (DMEM, F12, M9, LB and water) with a range of glucose spikes (Table 1). Mixtures of glucose and media were generated using the automated liquid handler of ambr®250 (Sartorius AG, Göttingen, Germany). A range of concentrated glucose stocks were prepared manually, their concentration verified by offline measurement using the Cedex®BioHT analyzer (Roche Custom Biotech, Penzberg, Germany) and added to a 24-deep well plate. 2.25 mL of each medium were dispensed into each well in a 48-deep-well plate, followed by the addition of 0.25 mL of glucose stock solution, resulting in 16 glucose concentrations between 0 and 20 g/L. Each medium was spiked with an identical set of glucose concentrations, and each medium–concentration combination was prepared in triplicate. Each mixture was mixed

by pipetting and measured using the at-line Raman analyzer (HyperFluxTM Raman Analyzer (Tornado, Mississauga, ON Canada)) integrated via the BioPAT@Spectra module (Sartorius AG, Göttingen, Germany). The Raman Analyzer is equipped with a 785 nm laser operating at a power of 495 mW. The spectral measurement range spans from 200 cm^{-1} to 3300 cm^{-1} (Raman shift) with a spectral resolution of $< 5 \text{ cm}^{-1}$. The acquisition settings were set to 1.2 s each and 30 accumulations were averaged for each sample. Acquisition settings were determined by adjusting settings based on medium with highest signal (TB, data not shown) that would result in 70% detector saturation. For each medium, the full sample set was prepared three times.

Dataset B (bioprocess scale-up)

Dataset B contains data from pilot (50 L and 150 L working volume) and production scale (14,000 L) main stage batches of a proprietary microbial (*Escherichia coli*) fed-batch fermentation for protein production (Table 1). The source and target domains therefore differ primarily in process scale, while the analyte ranges remain highly similar. The process is based on a chemically defined medium. During the main stage, carbon source and trace elements were continuously added. The process was operated under carbon-limited conditions, meaning that the carbon feed rate never exceeded the expected uptake by the microbial culture.

As a result, with the exception of the initial phase immediately after process initiation, glucose concentrations above 0.0 g/L should not occur. Transient glucose accumulation may occur under conditions of unintended overfeeding, for example when microbial metabolic activity declines. Accordingly, increasing glucose concentrations observed during the process may serve as an indicator of potential deviations from the intended process trajectory.

Raman spectra were collected inline using a Raman bIO-PRO probe (Kaiser Optical Systems, Ann Arbor, USA) installed in the bioreactor and connected to a Raman Rxn2 Analyzer (Endress+Hauser, Reinach, Switzerland). The analyzer was equipped with a 785 nm laser operated at 400 mW. Spectra were acquired over a Raman shift range of 100–3425 cm^{-1} with a spectral resolution of 1 cm^{-1} .

Each recorded spectrum was generated by averaging 60 consecutive individual acquisitions with an exposure time of 12.5 s each, resulting in a total integration time of 750 s per spectrum. Compared with the manufacturer-recommended setting of 10.0 s exposure time and 75 averaged acquisitions, the exposure time was increased to improve detector saturation, particularly during the early process phase. To keep the total integration time constant, the number of averaged acquisitions was reduced accordingly from 75 to 60. Spectral acquisition was performed periodically throughout the process, at least once per hour. Over the course of the process, detector saturation increased from approximately 10% to a maximum of 53%.

Offline analytics for reference samples was carried out with a SHIMADZU UV-1800 photometer for OD600 (Shimadzu, Kyoto, Japan) and with Thermo Scientific Arena 60 Analyzer for glucose and phosphate (Thermo Fisher, Massachusetts, USA). OD600 was used as a proxy for microbial biomass concentration.

Models

All modeling was performed using the following Python packages: chemotools [36], numpy [37], pandas [38], POT (Python Optimal Transport) [39], scikit-learn [40], pybaselines [41] and SKADA [42]. Six models were implemented, four of which leverage unlabeled target-domain spectra. For the models using target domain information and the reference model in Dataset A, standard normal variate (SNV) was consistently applied to the raw spectra as a preprocessing step. For the same models in Dataset B, data were preprocessed using a Savitzky–Golay first derivative filter followed by SNV. All model workflows are finalized with a Partial Least Squares (PLS) regression. The chosen model parameters can be found in Table S1 and Table S2 for Dataset A and B respectively.

Partial least squares regression (PLS) PLS is a regression technique that projects predictors and responses into a shared latent space to maximize their covariance, making it well-suited for high-dimensional, collinear data such as spectra [10, 43–45]. The built-in PLS implementation from scikit-learn was used, scanning 1–10 LVs. The optimal

Table 1 Overview of datasets and domains

Dataset	Use case	Source domain	Target domain	Analyte	N_S	N_T	B_S	B_T
A	Media transfer	DMEM, M9, F12, LB, or water	DMEM, M9, F12, LB, or water	Glucose	32	48 (if $S \neq T$); 16 (if $S = T$)	–	–
B	Scale transfer	Pilot scale	Production scale	Glucose, biomass, phosphate	1327	1427	15	22

N_S, N_T : number of spectra in source/target domain; B_S, B_T : number of batches in source/target domain

number of LVs was selected using a significance-based approach: Additional LVs were only included if they significantly reduced residuals (paired t-test, $p = 0.05$) [46]. To find the optimal number of LVs in CORAL, JDOT and TCA, a PLS model was trained on the source domain (5-fold CV; random splits). The found number of LVs was used for the PLS model after the domain adaptation step.

Optimal Preprocessing (OPP) The optimal preprocessing routine was based on the selection proposed by Engels et al. and is summarized in Table 2 [15]. For each preprocessing combination, a PLS model was trained on the source domain, and its performance was evaluated using 5-fold cross-validation (CV) with random splits. For all preprocessing combinations a significance test [46] was conducted to find the optimal number of latent variables (LVs) for a specific combination. The optimal model was chosen based on the lowest RMSE of cross-validation.

Base-case reference preprocessing (REF) For Dataset A, REF is a PLS model that is built on SNV corrected data. In Dataset B, a Savitzky–Golay filter (window size 20, first order derivative, third order polynomial) was used in conjunction with SNV normalization to enhance spectral feature resolution. These represent examples of basic preprocessing steps that are used when no systematic screening of preprocessing combinations is conducted [15].

Correlation Alignment (CORAL) [47–50]

CORAL is a domain adaptation method that aligns the second-order statistics (covariance matrices) of the source and target feature distributions (Fig. 1A). By minimizing the Frobenius norm between these covariances, CORAL transforms the source features to match the target distribution without requiring target labels. This simple, linear approach is computationally efficient and effective in reducing domain shift when variance structure is a key source of mismatch. In our workflow, Principal Component Analysis (PCA) was first applied to the combined source and target

spectra, retaining components that explained 99% of the total variance. CORAL was then performed in this reduced space, followed by PLS regression on the aligned features.

Joint Distribution Optimal Transport (JDOT) [51–53]

JDOT is a domain adaptation method that aligns both the feature and label distributions between source and target domains using optimal transport (Fig. 1B). It computes a coupling between source and target samples by minimizing a cost function that combines feature distances and label prediction errors, using pseudo-labels for the unlabeled target data. This joint alignment enables JDOT to transfer models more effectively when domain shifts affect both input features and label structures. The α parameter, which balances feature and label alignment, was selected as described by Courty et al. [51].

Optimal Preprocessing chosen by Maximum Mean Discrepancy (OPP-MMD)

This procedure was conducted based on Umprecht et al. [35]. In our implementation, we first retained the top 5% of preprocessing combinations according to the lowest RMSE on the source domain. For each corresponding PLS model, pseudo-labels were generated for the unlabeled target domain (Fig. 1C). The MMD between the source-domain labels and the target-domain pseudo-labels was then computed. The preprocessing combination yielding the lowest MMD was selected as the final optimal configuration. MMD is a statistical measure used to quantify domain differences. Here, it measures the domain distances between labels and pseudo-labels. A lower MMD indicates better alignment between domains, suggesting that the model generalizes more effectively to the target domain.

Transfer Component Analysis (TCA) [54–57]

TCA is a domain adaptation method that learns a shared feature space in which the marginal distributions of source and target domains are aligned (Fig. 1D). It achieves this by minimizing the Maximum Mean Discrepancy between

Table 2 Overview of applied preprocessing methods, grouped by correction type

Baseline	Scatter	Smoothing
None	None	Savitzky–Golay filter; no derivative and 1st order derivative window size: 10–60 (increments of 5) polynomial order: 2, 3, 4
1st–2nd order polynomial	Mean scaling	
AirPLS (10^5 – 10^{10} , multiplicative increase)	SNV	
AsLs (Asymmetric Least Squares) $\lambda : 10^4 - 10^7$; multiplicative increase $p : 0.1, 0.01$	Median scaling	
	Minmax scaling	
	Norm scaling	
	RNV	
	MSC	

SNV Standard normal variate, RNV Robust normal variate, MSC Multiplicative scatter correction, AirPLS Adaptive iteratively reweighted penalized least squares

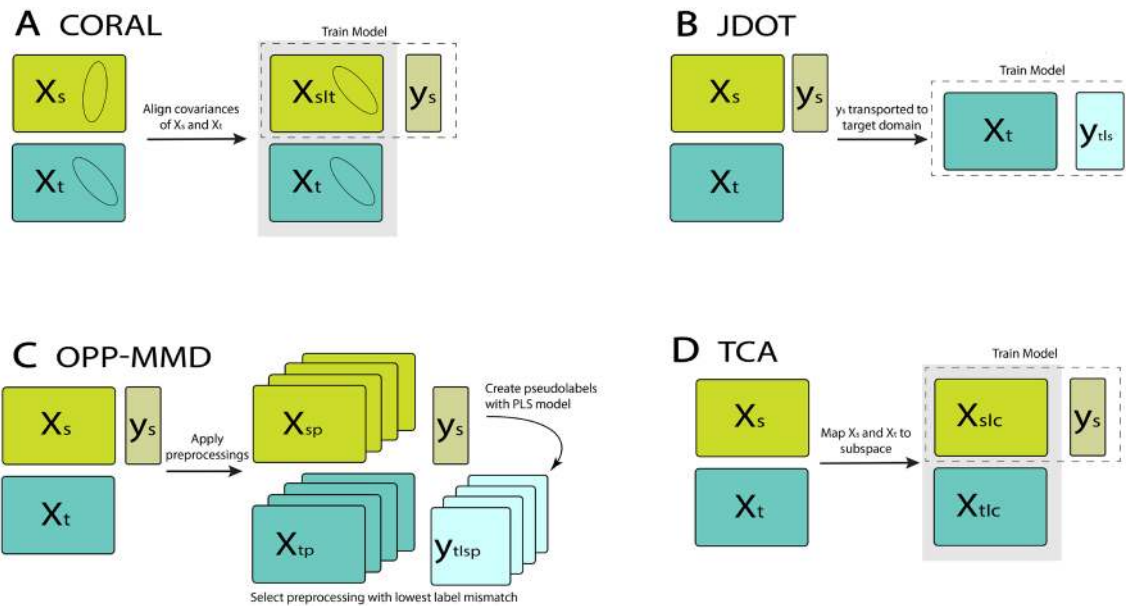


Fig. 1 Graphic representation of model working mechanisms; CORAL = Correlation alignment, JDOT = Joint distribution optimal transport, OPP-MMD = Optimal preprocessing chosen by maximum mean discrepancy, TCA = Transfer component analysis; X_s = source domain spectra; X_t = target domain spectra; y_s = source domain labels; $X_{s|t}$ = X_s with aligned covariances with target domain; $y_{t|s}$ = target

domain pseudo-labels created out of y_s ; X_{sp}/X_{tp} = X_s/X_t after pre-processing; $y_{t|sp}$ = preprocessed target domain pseudo-labels created out of y_s ; $X_{s|c}/X_{t|c}$ = X_s/X_t mapped to shared component space; $X_{t|c}$ = X_t mapped to subspace; grey background depicts a shared spectral domain; dashed rectangle depicts variables used for model training

domains while preserving the intrinsic structure of the data through kernel-based dimensionality reduction. Since TCA does not use label information, it is broadly applicable but may be less effective when task-relevant features differ across domains. TCA was applied to reduce the dimensionality of the spectra to 10 components and subsequently train a PLS model on this reduced space.

Performance metrics

In this study, R^2 , root mean squared error, and bias are used as complementary metrics to comprehensively evaluate model performance across domains. All used metrics are based on a comparison of the true values y_i and the predicted values $\hat{y}_i$ for all samples $i = 1, \dots, n$. Furthermore, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ refers to the average of the true values.

Coefficient of determination (R^2) The Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

quantifies the proportion of variance in the true analyte concentrations that is explained by the model, serving as a general indicator of goodness-of-fit.

Normalized Root Mean Squared Error (nRMSE)

In order to make the Root Mean Squared Error more comparable between domains and analytes with different scales we normalize it using the interquartile range (IQR) of the true values ($Q3 - Q1$), i.e.

$$nRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{Q3 - Q1} \quad (2)$$

where $Q1$ is the 25th percentile and $Q3$ is the 75th percentile of the dataset.

Normalized Bias (nBias) In the same way as for the nRMSE, we normalize the bias

$$nBias = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)}{Q3 - Q1} \quad (3)$$

This gives us a scale-free measure of the magnitude of systematic prediction errors, revealing whether models consistently over- or under-predict.

Maximum mean discrepancy (MMD)

MMD was used to quantify distributional differences between spectral datasets. MMD estimates the distance between two probability distributions and is zero only when the two distributions are identical. In this study, MMD was applied both to quantify inter- and intra-domain variability. For the latter, each domain was split into two subsets based on the mean analyte concentration within that domain.

Samples with analyte concentrations below the domain-specific mean were assigned to the “low” subset, whereas samples with concentrations above the mean were assigned to the “high” subset. The MMD between these two subsets was then used as a measure of intra-domain variability.

Noise simulations

For Dataset A, common sources of distortion in Raman spectroscopy were simulated to assess their impact on transfer learning performance. Specifically, additive white noise, peak shifts, baseline drifts, and scattering were considered, that can arise from low signal-to-noise ratio, inter-instrument variability, fluorescence, and bubbles as well as biomass respectively [58–61]. In addition, the effect of label distribution mismatch was examined, a factor of particular concern for unsupervised approaches, as it can naturally occur due to differences in sampling strategies, experimental design, or biological variation between source and target domains [23].

White Noise We simulated in-domain noisy data by adding a flat white noise to all spectra with standard deviation 300 [a.u.]. For comparison, the root mean squares of the spectrum amplitudes after first-order baseline correction lie between 1200 [a.u.] and 1400 [a.u.].

Peak Shifts Spectral peak positions were perturbed by applying small random displacements along the wavenumber axis, representing calibration inconsistencies between instruments. After application of second-order baseline correction and smoothing, we receive the spectral data $\tilde{S}(\nu)$. We then warp the spectrum by applying a displacement field to the wavelengths, thereby warping the coordinates $\nu' = \nu + \alpha \tilde{S}(\nu)$, shifting the peaks in the spectrum, where we randomly choose a different α for each of the domains (see Table 3).

Baseline Drift

To simulate complex between- and within-domain baseline shifts, we added a structured baseline $B(\nu)$ to each spectrum $X(\nu)$, where ν denotes the Raman shift [cm^{-1}]. Each culture medium has a unique baseline shape with per-sample stochastic parameters. We write $\mathcal{N}(0, \sigma^2)$ for a zero-mean Gaussian with standard deviation σ and use the logistic function $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$. Let $N_1 \sim \mathcal{N}(0, \sigma_1^2)$

and $N_2 \sim \mathcal{N}(0, \sigma_2^2)$ be sampled independently for each spectrum. The baselines are:

- *DMEM (linear downward tilt):*

$$B(\nu) = N_1 \nu + N_2, \quad N_1 \sim \mathcal{N}(0, 0.5^2), \\ N_2 \sim \mathcal{N}(0, 1000^2).$$

- *M9 (linear upward tilt):*

$$B(\nu) = (10 + N_1) \nu + N_2, \quad N_1 \sim \mathcal{N}(0, 0.5^2), \\ N_2 \sim \mathcal{N}(0, 1000^2).$$

- *LB (central peak; single-cycle sine):*

$$B(\nu) = (20000 + N_1) \sin(k \nu (1 + N_2)),$$

where $N_1 \sim \mathcal{N}(0, 500^2)$, $N_2 \sim \mathcal{N}(0, 0.025^2)$, and k is chosen so that $\sin(k \nu)$ completes exactly one half-period over the spectral range, yielding a single peak and zeros at the range endpoints.

- *F12 (sigmoid-shaped hump):*

$$B(\nu) = (20000 + N_1) \text{sigmoid}(k \nu (1 + N_2)),$$

where $N_1 \sim \mathcal{N}(0, 500^2)$, $N_2 \sim \mathcal{N}(0, 0.05^2)$, and k is the same value as in the LB condition.

- *H₂O (inverted sigmoid dip):*

$$B(\nu) = (-10000 + N_1) \text{sigmoid}(k \nu (1 + N_2)),$$

where $N_1 \sim \mathcal{N}(0, 500^2)$, $N_2 \sim \mathcal{N}(0, 0.05^2)$, and k matches the LB/F12 setting. **Scattering artifacts** For simulating the effects of bubbles and biomass we build on the insights of Klaverdijk et al. [61]. We used AirPLS ($\lambda = 10^5$) to isolate a baseline and then multiplied the remaining part by medium-dependent mixtures of a bubble (constant f^{Bubble}) and biomass (non-constant $f^{\text{Biomass}}(\nu)$) spectral variation interpolated from values given in [61]) contributions

$$\tilde{S}(\nu) = B^{\text{AirPLS}}(\nu) + (S(\nu) - B^{\text{AirPLS}}(\nu))(1 + c^{\text{Bubble}} f^{\text{Bubble}} + c^{\text{Biomass}} f^{\text{Biomass}}(\nu)). \quad (4)$$

The exact domain-specific contributions are summarized in Table 3.

Label Distribution Mismatch To induce uneven source–target label distributions, the source dataset was truncated by retaining only samples within a specified analyte range $[y_{\min}, y_{\max}]$. This procedure created systematic gaps in coverage relative to the target distribution.

Table 3 Noise simulation parameters by domain

Domain	α (approx.)	c^{Bubble}	c^{Biomass}
LB	0.0017	5.0	1.2
F12	−0.0088	2.5	0.6
M9	−0.0030	0.0	0.0
DMEM	0.0137	5.0	0.6
H ₂ O	−0.0198	0.0	0.6

Results

The results are organized in three steps. First, Dataset A is used to evaluate model behavior under controlled medium-transfer conditions, where the glucose concentration range is matched across domains. Second, in-silico perturbations and MMD analyses are used to identify when transfer learning improves over source-only calibration. Third, Dataset B tests whether the same behavior is observed in an industrial pilot-to-production scale-transfer setting. Across both datasets, we distinguish between cases where inter-domain spectral differences dominate and cases where analyte-related spectral variation is sufficiently preserved across domains.

Dataset A (media transfer)

Dataset A provides a controlled medium-transfer setting because all media were measured over the same glucose concentration range. Thus, differences in out-of-domain performance mainly reflect medium-induced spectral variation rather than differences in label coverage. Across all domain transfer scenarios, CORAL and JDOT consistently outperformed the other approaches in terms of predictive accuracy and generalization. Figure 2a shows that, for both methods, R^2 is equal to the in-domain training–testing baseline across all domain pairs (rounded to 1.00 across all domain pairs), indicating near-perfect agreement between predicted and true concentrations. Correspondingly, in Fig. 2b CORAL and JDOT achieved minimal systematic bias ($nBias \leq |1.6|\%$) and Fig. 2c further confirms very low prediction error ($nRMSE \leq 4.7\%$). Overall, JDOT and CORAL achieved performance comparable to source-domain cross-validation in Dataset A (Fig. S1).

OPP, OPP-MMD, REF, and TCA yielded comparable performance metrics across most domain combinations; however, OPP and OPP-MMD performance varied markedly when M9 or LB were involved (R^2 between -37.08 and 1.00 , absolute bias between 372.0% and 2.8% and RMSE between 372.4% and 3.7% ;), whereas REF and TCA performance declined when LB was involved (R^2 between -17.95 and 0.92 , absolute bias between 261.2% and 13.1% and nRMSE between 262.6% and 16.9% ;).

To gain deeper insight into conditions underlying model success or failure, we introduced artificial distortions representative of common Raman artifacts: white noise, peak shifts, baseline drifts, and scattering artifacts, typically arising from low signal-to-noise ratio, inter-instrument variability, fluorescence, bubbles, and biomass, respectively [58–61]. In Fig. 3, R^2 values of the original data and all in-silico simulations of OPP, CORAL and JDOT are shown. Normalized RMSE and normalized bias showed comparable qualitative trends to the original, unperturbed

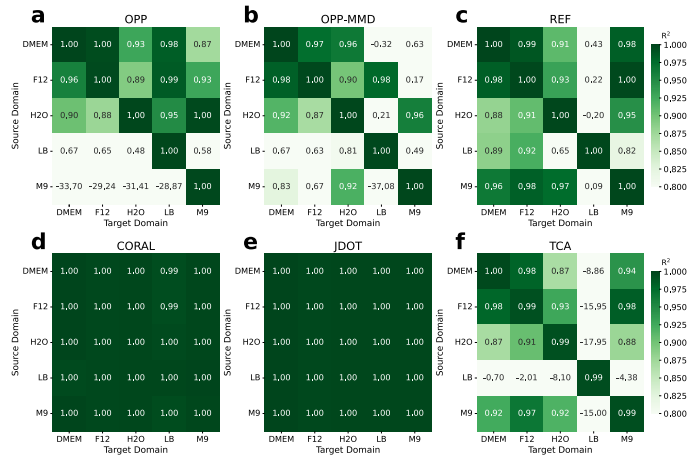
data. The results of the remaining models are provided in Figs. S2 to S5.

Under the influence of white noise (Fig. 3b), OPP performance improved in some transfer scenarios and decreased in others. Initially JDOT and CORAL performed worse in this setting due to a degradation of the base model, but when combined with a smoothing step, both TL methods recovered to levels almost that of the domains without the added noise (Fig. 3c). Baseline distortions of varying intensity and shape (linear, sigmoid, sinusoidal) highlighted JDOT as the most robust and consistent approach (Fig. 3d). OPP occasionally matched its accuracy but exhibited substantial variability across domain combinations. When spectral peaks were shifted, JDOT again outperformed all other methods (Fig. 3e), followed by CORAL, while the other models failed to generalize outside a given source domain. The simulation of scattering artifacts showed a trend consistent with the other evaluated noise types, with JDOT performing best, followed by CORAL (Fig. 3f). Finally, we examined the impact of mismatched label distributions between source and target domains, a frequent challenge in unsupervised transfer caused by differences in sampling or biological variation [23]. Focusing on JDOT, the overall best-performing model, we examined transfer from DMEM to F12. In this scenario, JDOT's performance declined under both mean and variance shifts (Fig. S10).

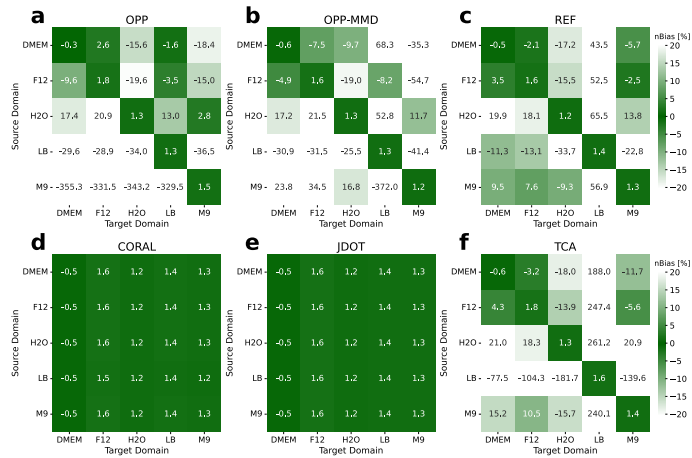
The MMD analysis was used to compare glucose-induced and medium-induced spectral variation. Spectra from each medium were split into low- and high-glucose subsets using the mean glucose concentration (8.8 g/L). Within-medium low–high MMD values reflect glucose-related variation, whereas between-medium MMD values at comparable glucose levels reflect medium-related domain variation. If the former is much larger, source-only transfer is expected to be sufficient; if both are comparable, medium effects are relevant and transfer learning can provide an advantage. In Dataset A, within-medium and between-medium MMD values were of comparable magnitude, indicating that medium-induced variation was a relevant source of spectral mismatch. This is illustrated in Fig. 4a, where the MMD values from “DMEM low” to “DMEM high”, “F12 low”, and “F12 high” are similar. A subset of the corresponding MMD matrix is shown in Fig. 4b, and the complete matrix is provided in Fig. S12.

The regression coefficients of the evaluated models can be found in Fig. S13a and show distinct glucose-related Raman bands around 1125 cm^{-1} and 517 cm^{-1} [62]. For TCA, regression coefficients cannot be shown since this model is fitted after projection into a latent space.

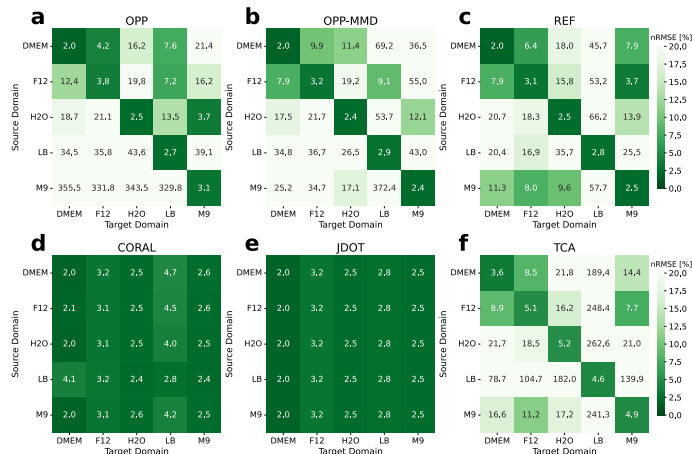
Fig. 2 Model performance in Dataset A across multiple medium-to-medium transfer scenarios. Each subfigure compares six modeling approaches when trained on one medium and evaluated on the remaining media. **a** OPP = optimal preprocessing, **b** OPP-MMD = optimal preprocessing selected based on maximum mean discrepancy, **c** REF = reference preprocessing, **d** CORAL = correlation alignment, **e** JDOT = joint distribution optimal transport, and **f** TCA = transfer component analysis



(a) R^2 . Higher values indicate better agreement between predicted and measured glucose concentrations.



(b) Normalized prediction bias (nBias [%]). Values close to zero indicate low systematic over- or under-prediction.



(c) Normalized root mean squared error (nRMSE [%]). Lower values indicate smaller prediction errors.

Fig. 3 R^2 values for Dataset A across different domain-transfer scenarios and simulated Raman noise conditions. Columns show the three evaluated modeling approaches: OPP (optimized pre-processing), CORAL (correlation alignment), and JDOT (joint distribution optimal transport). Rows correspond to the individual noise scenarios (a–f). Each heatmap shows model performance when trained on one source domain and tested on the remaining target domains. Higher R^2 values indicate better agreement between predicted and measured glucose concentrations. The OPP model selected for **b** White noise and **c** White noise—smooth was identical; therefore, the corresponding heatmap is shown only once

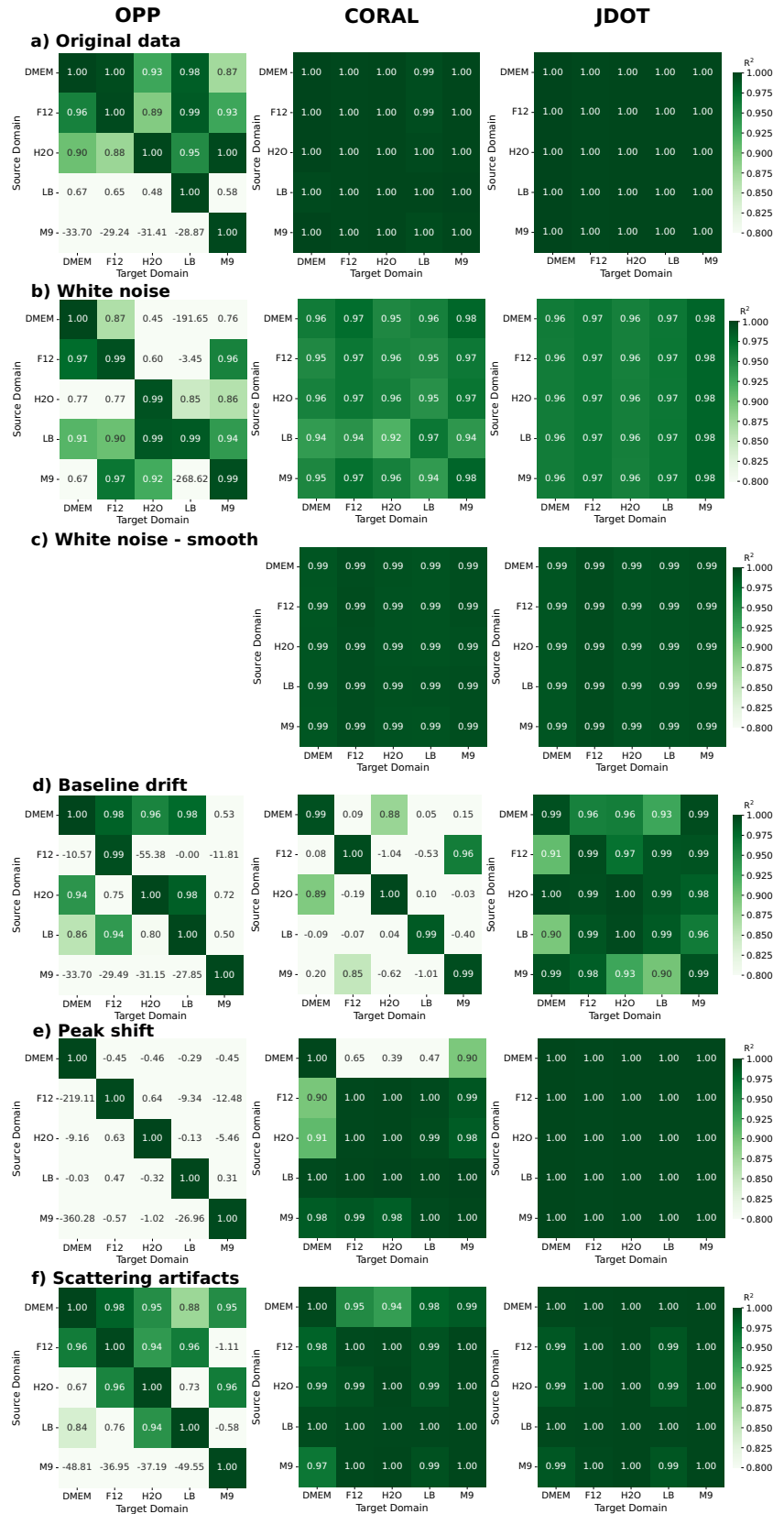
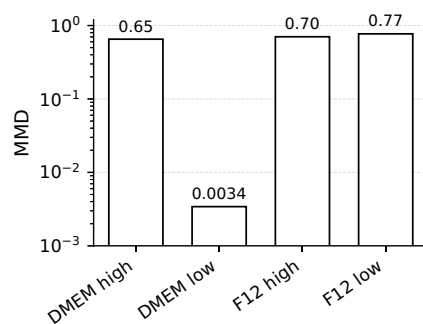
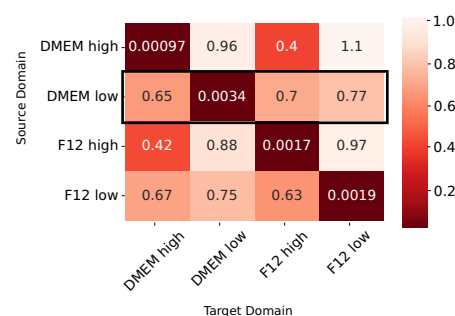


Fig. 4 Examples of the MMD-based comparison of spectral subset variability in Dataset A

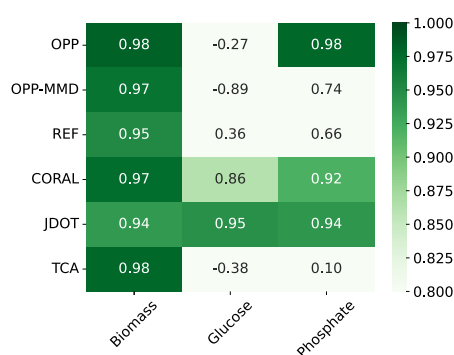


(a) MMD values between the glucose reference subset, “DMEM low”, and the indicated DMEM and F12 concentration subsets.

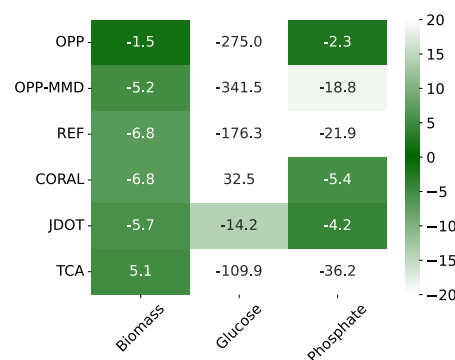


(b) Extract of split-domain differences quantified using MMD (maximum mean discrepancy). Lower MMD values indicate greater similarity between the compared domains.

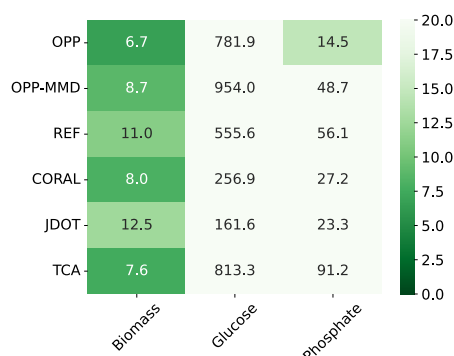
Fig. 5 Performance metrics for Dataset B across six modeling techniques applied to a bioprocess scale-up scenario. Heatmap (a)–(c) summarizes one metric— R^2 , nBias [%] or nRMSE [%]—across all analytes. R^2 = coefficient of determination, nBias [%] = normalized bias in percent, nRMSE [%] = normalized root mean squared error in percent, OPP = Optimal preprocessing, OPP-MMD = Optimal preprocessing chosen by maximum mean discrepancy, REF = Reference preprocessing, CORAL = Correlation alignment, JDOT = Joint distribution optimal transport, TCA = Transfer component analysis; each model workflow contains a preprocessing and/or transfer step followed by a PLS regression



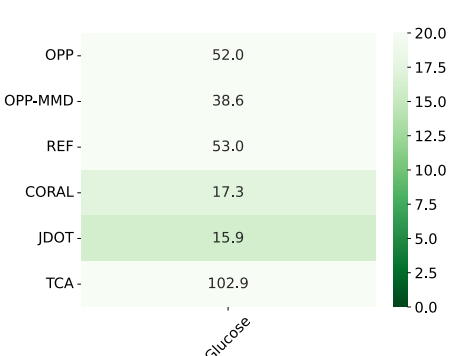
(a) R^2



(b) nBias [%]



(c) nRMSE [%]



(d) RMSE/mean [%] for elevated glucose values.

Dataset B (bioprocess scale-up)

Dataset B evaluated model transfer from pilot- to production-scale fermentations for three analytes: biomass, glucose, and phosphate. Model performance is summarized in Fig. 5 and differed between analytes.

For biomass, all evaluated models showed comparably strong performance with $R^2 \geq 0.94$, $|\text{nBias}| \leq 6.8\%$, and $\text{nRMSE} \leq 12.5\%$. OPP achieved the lowest prediction

error for biomass, with $R^2 = 0.98$, $\text{nBias} = -1.5\%$, and $\text{nRMSE} = 6.7\%$.

For glucose, the concentration distribution was strongly skewed, with many samples close to zero and fewer samples at elevated concentrations (Fig. S14). Therefore, global IQR-normalized errors were strongly influenced by the narrow interquartile range of the glucose distribution. Across all glucose samples, JDOT followed by CORAL achieved the lowest prediction error among the evaluated models,

with $R^2 = 0.95$, $|nBias| = 14.2\%$, and $nRMSE = 161.6\%$ for JDOT, and $R^2 = 0.86$, $|nBias| = 32.5\%$, and $nRMSE = 256.9\%$ for CORAL. The source-only reference models showed larger errors and stronger prediction bias. Because the global glucose metrics were strongly affected by the narrow interquartile range of the skewed concentration distribution, prediction error was additionally calculated for the elevated-glucose samples only (Fig. 5d). In this subset, JDOT and CORAL showed RMSE values of 15.9% and 17.3% compared to the mean of the subset, respectively, whereas the source-only reference models OPP and REF showed RMSE values above 50%.

For phosphate, OPP, JDOT, and CORAL showed the highest performance, with $R^2 \geq 0.92$, $|nBias| \leq 5.4\%$, and $nRMSE \leq 27.2\%$. Across analytes, JDOT showed consistently high model performance, with R^2 values above 0.90 for biomass, glucose, and phosphate. TCA and OPP-MMD did not improve over the source-only reference models. The regression coefficients for the analytes and models used on Dataset B are shown in Figs. S13b to S13d.

The MMD analysis indicated analyte-dependent spectral variability patterns in the pilot-to-production transfer setting (Fig. 6).

Discussion

This study set out to address the question of whether unsupervised transfer learning can enhance out-of-domain model performance in settings relevant for upstream bioprocessing. Such situations commonly occur, for example when a model calibrated in one cultivation medium is used in another medium, or when a model developed at small scale is transferred to pilot or production scale. We therefore evaluated transfer learning methods on both controlled spiked-media datasets and industrially relevant scale-up datasets. Overall, the results show that transfer learning did not universally surpass the benchmarks. However, selected transfer learning methods improved prediction performance when the spectral differences between source and target domains were larger than the variation within each domain condition.

Dataset A provided a well-controlled testbed: glucose was the analyte of interest, and different domains were created by switching between growth media. The same glucose distribution was present in all domains, such that the transfer task consisted of correcting medium-induced spectral differences. This dataset structure enabled a systematic

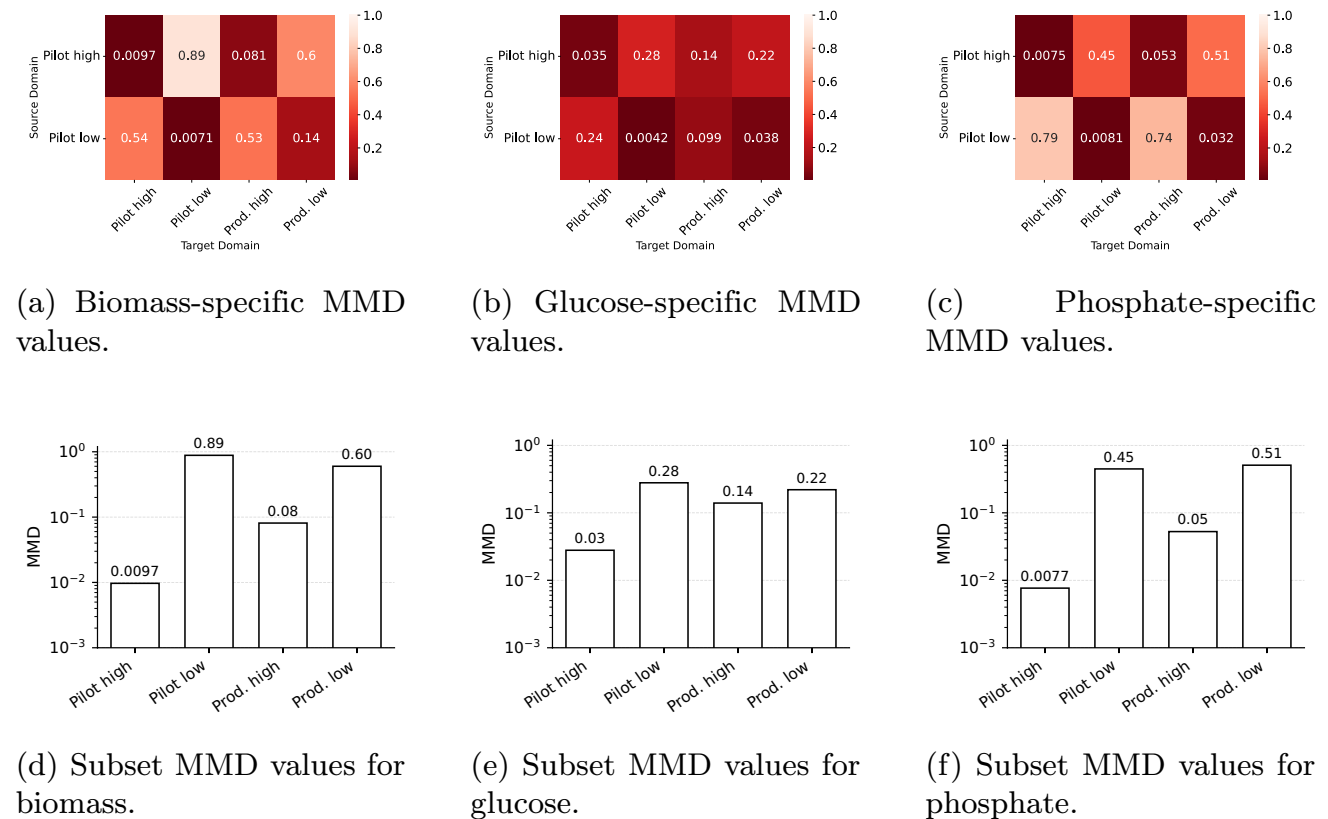


Fig. 6 MMD-based analysis of spectral variability in Dataset B. Top row: analyte-specific MMD comparisons across pilot- and production-scale spectra for biomass, glucose, and phosphate. Lower MMD values indicate greater similarity between the compared spectral distributions.

Bottom row: MMD values between the respective reference subset and the indicated pilot- or production-scale concentration subsets. For all analytes, the reference subset was “Pilot high”

comparison of model behavior under domain variability. The results show that JDOT and, except for simulated baseline shifts, also CORAL consistently outperformed the baseline model OPP, which was built on source-domain data alone (Fig. 3).

The MMD analysis helps to explain why these transfer learning approaches were beneficial in Dataset A. The comparison between low- and high-glucose groups in Fig. 4 showed that varying glucose concentrations as well as domain switches to another medium caused clear spectral changes. This effect is depicted in Fig. 4a: The MMD values between “DMEM low” and the other domain subgroups, “DMEM high”, “F12 high” and “F12 low”, are of comparable magnitude with 0.65, 0.70 and 0.77 respectively. This indicates that changing the medium introduced spectral variation that was similarly strong as the variation caused by glucose concentration within one medium. This observation is important for the interpretation of the model results. A model trained only in one medium may not only learn glucose-related spectral features, but also medium-specific spectral structures. When this model is then applied to another medium, these medium-related differences can reduce prediction performance. This provides a likely explanation for the weaker performance of the source-domain-based reference models OPP and REF in Dataset A. JDOT and CORAL are designed to reduce such differences between source and target spectra. CORAL aligns the covariance structure of the source and target data, while JDOT aligns samples by considering both spectral similarity and predicted label structure. Therefore, both methods can reduce medium-related spectral mismatch. The MMD results support this interpretation, because they show that medium effects were a relevant source of spectral variation in Dataset A.

To verify that TL methods perform beyond idealized settings, we introduced additional perturbations and noise types representative of Raman measurements. Overall, JDOT showed the best performance, and the gap to source-domain-based models (OPP, REF) tended to widen under perturbations. This is most apparent under peak shifts, which are common Raman artifacts arising among others from instrument-to-instrument variability [58]. JDOT largely mitigated systematic spectral misalignment, reaching accuracy close to a fully supervised PLS model trained and tested in the target domain (Fig. 3e). Its behavior is consistent with its optimal-transport alignment in joint feature-label space, which can mitigate moderate distribution shifts. CORAL also performed strongly under peak shifts, underscoring the effectiveness of fast, covariance-alignment approaches. In contrast, source-domain-based models deteriorated markedly, lacking a mechanism to handle shifted peaks.

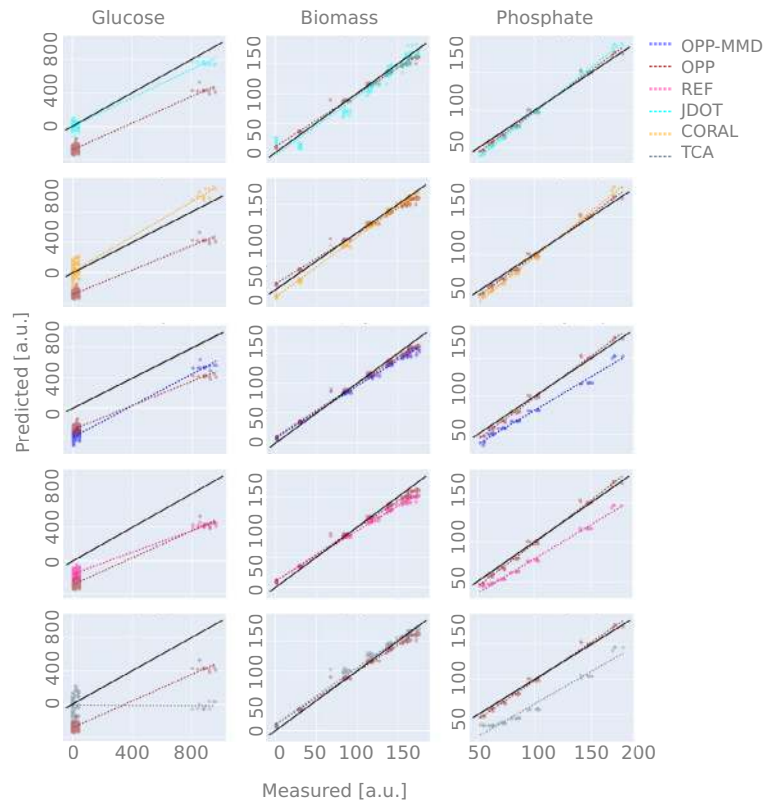
Scattering artifacts, a common bioreactor effect caused by factors such as bubbles and biomass, were handled most effectively by JDOT and, for most domain combinations, also by CORAL (Fig. 3f). Under baseline drift across and within domains, JDOT showed the strongest performance and clearly exceeded all other methods, consistent with its label transport strategy (Fig. 3d). Overall, these results indicate that among the tested methods, JDOT is the least sensitive to several noise types commonly encountered in Raman spectroscopy for bioprocess monitoring.

When considering the presence of additional white noise, results highlight that preprocessing and TL (particularly CORAL and JDOT) act as complementary rather than competing components. Taking JDOT as an example, while JDOT manages to assign accurate pseudo-labels to the target domain data, the performance bottleneck is the base PLS model due to in-domain noise. By applying smoothing to the spectra, the PLS model performance can be improved, and near-optimal performance for JDOT and CORAL is regained (Fig. 3b and c).

We also observed a practical limitation of unsupervised transfer learning. When the glucose distribution in the target medium differed from the source medium (Fig. S10), JDOT performance decreased. Importantly, this was the only evaluated scenario in which the transfer task moved from interpolation toward extrapolation, because the target-domain analyte distribution was no longer fully covered by the source-domain calibration range. In line with literature, this indicates that unsupervised TL is sensitive to differences in analyte distributions between source and target domains [23]. Feature-based alignment methods such as CORAL and TCA face similar challenges, because they align spectral feature distributions but do not explicitly correct differences in the analyte distribution between source and target domains. Semi-supervised approaches, for example semi-supervised JDOT with a small number of labeled target samples, may reduce this limitation in practical applications.

If Dataset A showed the behavior of the transfer-learning methods under controlled medium shifts, Dataset B tested whether similar effects could also be observed in an industrially relevant bacterial bioprocess scale-transfer setting. The results were strongly analyte-dependent. For biomass and phosphate, where the MMD analysis indicated that analyte-related variation was pronounced (Fig. 6), the optimized PLS-based reference workflow performed competitively or better than the tested transfer-learning methods. This suggests that, when the analyte signal is sufficiently preserved across scales, careful preprocessing and PLS model optimization can already compensate for a substantial part of the domain difference.

Fig. 7 Measured-vs-predicted plots of analytes modeled in Dataset B. Data points were normalized according to the mean of a given analyte range. OPP = Optimal preprocessing, MMD = Maximum mean discrepancy, REF = Reference preprocessing, CORAL = Correlation alignment, JDOT = Joint distribution optimal transport, TCA = Transfer component analysis; each model workflow contains a preprocessing and/or transfer step followed by a PLS regression



Glucose showed a different behavior and has to be interpreted in the context of the glucose-limited process operation. After the initial process phase, glucose concentrations are expected to remain close to zero, whereas elevated glucose values indicate transient deviations from the intended carbon-limited state, for example due to overfeeding or reduced uptake. This process behavior is reflected in the strongly skewed glucose distribution in Dataset B, with many near-zero samples and only a smaller elevated-glucose subset (Fig. S14). As a result, global normalized error metrics are difficult to interpret for glucose, because the interquartile range is dominated by the near-zero concentration regime and small absolute errors can lead to large normalized errors. Therefore, the glucose results should not be interpreted as a broad quantitative calibration over a uniformly sampled concentration range, but in the context of a glucose-limited process with a dominant near-zero state and a smaller elevated-glucose subset.

For this process, the more relevant monitoring task is to distinguish glucose-depleted from elevated-glucose states and to recover the elevated-glucose level more accurately. In this elevated-glucose subset, JDOT and CORAL achieved lower errors than the source-only workflows, with RMSE values of 15.9% and 17.3% of the mean value, respectively, whereas OPP and REF showed errors above 50% (Fig. 5d). This indicates that the transfer-learning methods predicted the elevated-glucose subset closer to the correct

concentration level on average and improved the prediction behavior in the process-relevant glucose states.

The improvements of the transfer-learning methods can be understood by considering the measured-versus-predicted plots. These plots suggest that the improved prediction behavior was largely related to a reduction of the scale-dependent prediction offset (Fig. 7). The source-only models showed negative intercept shifts after transfer from pilot to production scale, whereas JDOT and CORAL showed intercepts closer to zero. This is consistent with the transfer mechanisms of the two methods: CORAL aligns second-order spectral statistics between pilot- and production-scale spectra, whereas JDOT additionally considers the joint structure of spectral features and predicted labels during transport.

Overall, Dataset B is consistent with the interpretation derived from Dataset A: unsupervised transfer learning is most useful when inter-domain spectral variability is large relative to analyte-induced variation. The glucose case illustrates this behavior in a challenging, glucose-limited process, where JDOT and CORAL mainly improved scale-transfer behavior by reducing the prediction offset and recovering the elevated-glucose state more accurately than the source-only workflows.

Across both datasets, the comparison between REF and OPP also showed that preprocessing selection remained an important determinant of model performance. For biomass

and phosphate in Dataset B, OPP performed competitively or better than the tested transfer-learning methods, indicating that systematic preprocessing selection and PLS model optimization can already compensate for a substantial part of the domain difference when analyte-related spectral information is sufficiently preserved across scales. This also shows that transfer learning should not be viewed as a replacement for careful chemometric model development, but as a complementary strategy for cases where preprocessing and source-domain calibration alone are insufficient.

Overall, among the tested transfer-learning methods, JDOT and CORAL showed the most consistent benefits. CORAL reduces domain differences by aligning second-order spectral statistics between source and target spectra, whereas JDOT uses optimal transport to align source and target samples while also considering the predicted label structure. This likely explains why both methods were most beneficial in cases where scale- or medium-related spectral variation was large relative to the analyte-induced variation. In contrast, TCA, which had previously shown strong performance on IR spectra, failed here, perhaps due to Raman- and domain-specific factors [30]. This underscores that TL methods do not generalize uniformly across spectral modalities and therefore require careful method selection, tuning, and validation before deployment in upstream bioprocess applications.

From an industrial perspective, the intended use case of the best-performing transfer-learning method is to reduce the recalibration effort when moving to a related target process, such as a new medium, scale, or reactor. In such a workflow, a calibration model developed on the source process is adapted using unlabeled target spectra acquired during early runs, after which the adapted model can be used for on-line prediction. In practice, deployment would still require standard PAT workflows, including instrument qualification, a modeling pipeline, predefined acceptance limits, and ongoing model maintenance.

In this context, JDOT stands out not just for its solid performance under common noise types encountered in Raman spectroscopy during bioprocesses but also for its efficiency and ease of use. It required under a minute to process all domain pairs in Dataset A, required no labeled target data, and could be implemented directly in Python without extensive tuning. In contrast, OPP-based workflows were computationally expensive, often exceeding 30 min per analyte due to preprocessing selection and MMD evaluation. Importantly, JDOT was consistently competitive with the baselines REF and OPP and superior when domain shifts dominated, making it a practically attractive candidate for improving generalization when labeled target data are not yet available.

Still, this study has limitations. We used only PLS regression, which may bias results toward linear models. Our preprocessing pipeline for the transfer-learning workflows was fixed and not further optimized per analyte, as we focused on non-expert-driven choices; however, expert knowledge and more flexible customization of preprocessing—especially in combination with JDOT and CORAL—could further enhance results. As seen in this study, preprocessing and transfer learning are complementary, and their combination may yield additional performance gains. Also, the performance of OPP-MMD could likely be improved by choosing absolute thresholds adjusted and optimized per analyte, instead of a relative, uniform threshold. Moreover, we excluded semi-supervised approaches, even though a small number of labeled target samples could, in theory, help methods like JDOT resolve label alignment ambiguity. These constraints could affect method rankings and will need to be addressed in follow-up work.

Taken together, the results indicate that unsupervised transfer learning, particularly JDOT and CORAL, can improve Raman spectroscopy-based model transfer when domain-related spectral variation limits source-only calibration. However, the benefit depends on the relation between analyte-induced spectral variation and inter-domain spectral differences. When analyte-related variation is sufficiently preserved across domains, optimized preprocessing and PLS modeling can remain competitive. When domain differences dominate, unsupervised transfer learning provides a useful additional strategy for improving scale- or medium-transfer performance without requiring labeled target-domain samples during model development.

Conclusion

This work evaluated four unsupervised transfer learning methods—CORAL, JDOT, OPP-MMD, and TCA—for predicting analyte concentrations from Raman spectra under both standardized media (Dataset A) and industrial scale-up (Dataset B) conditions. Our results highlight that JDOT, and in several cases also CORAL, offers a practical complement to preprocessing-driven modeling approaches. It was competitive with or outperformed OPP in several transfer scenarios and does not require extensive expert-guided preprocessing selection, making it attractive for bioprocess modeling. However, the advantage of JDOT was conditional and depended on the relation between domain-induced spectral variation and analyte-induced spectral variation. When the analyte contribution to domain variability was large, non-transfer alternatives performed comparably or better, suggesting that the benefits of TL are context-specific. Furthermore, the study was limited to fixed

preprocessing routines, excluded semi-supervised alternatives, and relied solely on PLS regression as the predictive model. These constraints may have influenced the observed model rankings and should be addressed in future work.

Overall, especially JDOT represents a promising and computationally efficient domain adaptation in upstream bioprocessing, but the superiority is not guaranteed. Therefore, future work should focus on systematically investigating the influence of semi-supervised models and non-linear regression models, particularly in industrial-scale bioprocesses.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00449-026-03394-8>.

Acknowledgements All authors acknowledge funding by the FFG project Interpretable and Interactive Transfer Learning in Process Analytical Technology (INTRAL, Grant No. 883856). Baxalta Innovations GmbH, a member of the Takeda group of companies has partially funded the study through provision of resources for data generation. Novartis Pharmaceutical Manufacturing GmbH has partially funded the study through provision of resources for data generation. We thank Valeria Fonseca-Diaz for fruitful discussions.

Author contributions BH wrote the original draft, reviewed and edited the manuscript, prepared visualizations, developed software and methodology, and conducted the investigation. AU wrote the original draft, reviewed and edited the manuscript, performed experimental work, developed software, and contributed to methodology. PE wrote the original draft and reviewed and edited the manuscript. BK reviewed and edited the manuscript. OS reviewed and edited the manuscript and provided resources.

Funding Open access funding provided by TU Wien (TUW). Funding by the FFG project Interpretable and Interactive Transfer Learning in Process Analytical Technology (INTRAL, Grant No. 883856). Baxalta Innovations GmbH, a member of the Takeda group of companies, and Novartis Pharmaceutical Manufacturing GmbH have partially funded the study through provision of resources for data generation.

Data availability The data used in this study are confidential.

Declarations

Conflict of interest Bianca Hüpfl is an employee of SEPA Engineering GmbH and is seconded to Baxalta Innovations GmbH, a member of the Takeda group of companies, and may own company stock. Alexandra Umrecht is an employee of Baxalta Innovations GmbH, a member of the Takeda group of companies, and may own company stock. Peter Ettinger is an employee of Novartis Pharmaceutical Manufacturing GmbH and may own company stock. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. FDA: Pat-a framework for innovative pharmaceutical development, manufacturing and quality assurance. US Food and Drug Administration (2004)
2. FDA: Pharmaceutical cgmcs for the 21st century—a risk-based approach. US Food and Drug Administration (2004)
3. Rathore AS, Winkle H (2009) Quality by design for biopharmaceuticals. *Nat Biotechnol* 27(1):26–34
4. Yu LX, Amidon G, Khan MA, Hoag SW, Polli J, Raju G, Woodcock J (2014) Understanding pharmaceutical quality by design. *AAPS J* 16:771–783
5. Gerzon G, Sheng Y, Kirkitadze M (2022) Process analytical technologies—advances in bioprocess integration and future perspectives. *J Pharm Biomed Anal* 207:114379
6. Esmonde-White KA, Cuellar M, Uerpmann C, Lenain B, Lewis IR (2017) Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Anal Bioanal Chem* 409(3):637–649
7. Lawrence XY, Kopcha M (2017) The future of pharmaceutical quality and the path to get there. *Int J Pharm* 528(1–2):354–359
8. O'Connor T, Chatterjee S, Lam J, Ossa DHP, Martinez-Peyrat L, Hoefnagel MH, Fisher AC (2024) An examination of process models and model risk frameworks for pharmaceutical manufacturing. *Int J Pharm X*, 100274
9. Destro F, Barolo M (2022) A review on the modernization of pharmaceutical development and manufacturing—trends, perspectives, and the role of mathematical modeling. *Int J Pharm* 620:121715
10. Wold S, Sjöström M, Eriksson L (2001) Pls-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58(2):109–130
11. Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359
12. Helleckes LM, Hemmerich J, Wiechert W, Lieres E, Grünberger A (2023) Machine learning in bioprocess development: from promise to practice. *Trends Biotechnol* 41(6):817–835
13. Bocklitz T, Walter A, Hartmann K, Rösch P, Popp J (2011) How to pre-process Raman spectra for reliable and stable models? *Anal Chim Acta* 704(1–2):47–56
14. Guo S, Popp J, Bocklitz T (2021) Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling. *Nat Protoc* 16(12):5426–5459
15. Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, Buydens LM (2013) Breaking with trends in pre-processing? TrAC, *Trends Anal Chem* 50:96–106
16. Troup GM, Georgakis C (2013) Process systems engineering tools in the pharmaceutical industry. *Comput Chem Eng* 51:157–171
17. Brunel B, Alsamad F, Piot O (2021) Toward automated machine learning in vibrational spectroscopy: use and settings of genetic algorithms for pre-processing and regression optimization. *Chemom Intell Lab Syst* 219:104444
18. Mishra P, Roger JM, Marini F, Biancolillo A, Rutledge DN (2022) Pre-processing ensembles with response oriented sequential alternation calibration (prosaac): a step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemom Intell Lab Syst* 222:104497

19. Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S (1998) An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom Intell Lab Syst* 44(1–2):229–244
20. Pétillet L, Pewny F, Wolf M, Sanchez C, Thomas F, Sarrazin J, Fauland K, Katinger H, Javalet C, Bonneville C (2020) Calibration transfer for bioprocess Raman monitoring using Kennard stone piecewise direct standardization and multivariate algorithms. *Eng Rep* 2(11):12230
21. Rowland-Jones RC, Graf A, Woodhams A, Diaz-Fernandez P, Warr S, Soeldner R, Finka G, Hoehse M (2021) Spectroscopy integration to miniature bioreactors and large scale production bioreactors-increasing current capabilities and model transfer. *Biotechnol Prog* 37(1):3074
22. Yang N, Guerin C, Kokanyan N, Perré P (2024) In-line monitoring of bioreactor by Raman spectroscopy: Direct use of a standard-based model through cell-scattering correction. *J Biotechnol* 396:41–52
23. Nikzad-Langerodi R, Andries E (2021) A chemometrician's guide to transfer learning. *J Chemom* 35(11):3373
24. Roger J-M, Diaz VF, Nikzad-Langerodi R (2025) Cloning instruments, model maintenance and calibration transfer. *TrAC, Trends Anal Chem* 191:118319
25. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J (2020) Transfer learning for drug discovery. *J Med Chem* 63(16):8683–8694
26. Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei G-W (2023) Machine learning methods for small data challenges in molecular science. *Chem Rev* 123(13):8736–8780
27. Chai Z, Zhao C, Huang B (2025) Cross-domain knowledge transfer in industrial process monitoring: a survey. *J Process Control* 149:103408
28. Padarian J, Minasny B, McBratney A (2019) Transfer learning to localise a continental soil vis-nir calibration model. *Geoderma* 340:279–288
29. Hu J, Zou Y, Sun B, Yu X, Shang Z, Huang J, Jin S, Liang P (2022) Raman spectrum classification based on transfer learning by a convolutional neural network: application to pesticide detection. *Spectrochim Acta Part A Mol Biomol Spectrosc* 265:120366
30. Nikzad-Langerodi R, Zellinger W, Saminger-Platz S, Moser BA (2020) Domain adaptation for regression under Beer-Lambert's law. *Knowl-Based Syst* 210:106447
31. Zhou W, Tang Y, Qian Z, Wang J, Guo H (2022) Deeply-recursive convolutional neural network for Raman spectra identification. *RSC Adv* 12(8):5053–5061
32. Huang T-Y, Yu JCC (2021) Development of crime scene intelligence using a hand-held Raman spectrometer and transfer learning. *Anal Chem* 93(25):8889–8896
33. Shi Z, Wang J, Su Y, Liang X, Zi J, Wang C, Bi H, Xiang X (2025) Transfer contrastive learning for Raman spectra data of urine: detection of glucose, protein, and prediction of kidney disorders. *Chemom Intell Lab Syst* 261:105384
34. Liu Y, Chen C, Zuo E, Yan Z, Chang C, Cheng Z, Lv X, Chen C (2024) Murda: multisource unsupervised Raman spectroscopy domain adaptation model with reconstructed target domains for medical diagnosis assistance. *Anal Chem* 96(39):15540–15549
35. Umprrecht A, Diaz VF, Mikulasek B, Kozma B, Schwaighofer A, Henson M, Nikzad-Langerodi R, Spadiut O (2025) Unsupervised optimization of spectral pre-processing selection to achieve transfer of Raman calibration models. *Measurement* 117906
36. Lopez PC (2024) chemotools: a python package that integrates chemometrics and scikit-learn. *J Open Source Softw* 9(100):6802. <https://doi.org/10.21105/joss.06802>
37. Harris CR, Millman KJ, Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, Kerkwijk MH, Brett M, Haldane A, Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. *Nature* 585(7825):357–362. <https://doi.org/10.1038/s41586-020-2649-2>
38. McKinney: data structures for statistical computing in python. In: Walt, Millman (eds) *Proceedings of the 9th python in science conference*, pp 56–61 (2010). <https://doi.org/10.25080/Majora-92bf1922-00a>
39. Flamary R, Vincent-Cuaz C, Courty N, Gramfort A, Kachaiev O, Quang Tran H, David L, Bonet C, Cassereau N, Gnassounou T, Tanguy E, Delon J, Collas A, Mazelet S, Chapel L, Kerdoncuff T, Yu X, Feickert M, Krzakala P, Liu T, Fernandes Montesuma E (2024) POT python optimal transport (version 0.9.5). <https://github.com/PythonOT/POT>
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
41. Erb D.: pybaselines: a python library of algorithms for the baseline correction of experimental data. <https://doi.org/10.5281/zenodo.5608581>. <https://github.com/derb12/pybaselines>
42. Gnassounou T, Kachaiev O, Flamary R, Collas A, Lalou Y, Mathelin A, Gramfort A, Bueno R, Michel F, Mellot A, Loison V, Odonnat A, Moreau T (2024) SKADA: Scikit adaptation. <https://scikit-adaptation.github.io/>
43. Wold H (1982) Soft modeling: the basic design and some extensions. *Systems under indirect observation. Part II*(2):36–37
44. Nicolai BM, Beullens K, Bobelyn E, Peirs A, Saey W, Theron KI, Lammertyn J (2007) Nondestructive measurement of fruit and vegetable quality by means of nir spectroscopy: a review. *Postharvest Biol Technol* 46(2):99–118
45. Morais CL, Lima KM, Singh M, Martin FL (2020) Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat Protoc* 15(7):2143–2162
46. Mikulasek B, Fonseca Diaz V, Gabauer D, Herwig C, Nikzad-Langerodi R (2023) Partial least squares regression with multiple domains. *J Chemom* 37(5):3477
47. Sun B, Feng J, Saenko K (2016) Return of frustratingly easy domain adaptation. In: *Proceedings of the AAAI conference on artificial intelligence*, vol 30
48. Feng L, Wu B, He Y, Zhang C (2021) Hyperspectral imaging combined with deep transfer learning for rice disease detection. *Front Plant Sci* 12:693521
49. Zhong X-C, Wang Q, Liu D, Liao J-X, Yang R, Duan S, Ding G, Sun J (2023) A deep domain adaptation framework with correlation alignment for eeg-based motor imagery classification. *Comput Biol Med* 163:107235
50. Shao X, Kim C-S (2022) Unsupervised domain adaptive 1d-cnn for fault diagnosis of bearing. *Sensors* 22(11):4156
51. Courty N, Flamary R, Habrard A, Rakotomamonjy A (2017) Joint distribution optimal transportation for domain adaptation. *Adv Neural Inform Process Syst*, p 30
52. Ackaouy A, Courty N, Vallée E, Commowick O, Barillot C, Galassi F (2020) Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data. *Front Comput Neurosci* 14:19
53. Damodaran BB, Kellenberger B, Flamary R, Tuia D, Courty N (2018) Deepjdot: deep joint distribution optimal transport for unsupervised domain adaptation. In: *Proceedings of the European conference on computer vision (ECCV)*. pp 447–463
54. Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. *IEEE Trans Neural Netw* 22(2):199–210

55. Zhang Z, Liu H, Wei Z, Lu M, Pu Y, Pan L, Zhang Z, Zhao J, Hu J (2023) A transfer learning method for spectral model of moldy apples from different origins. *Food Control* 150:109731
56. Zhu X, Liu W, Wang B, Wang W (2022) A soft sensor model of pichia pastoris cell concentration based on ibda-reln. *Prep Biochem Biotechnol* 52(6):618–626
57. Cheng Q, He B, Zhao C, Bi H, Chen D, Han S, Gao H, Feng W (2021) Prediction of functional microexons by transfer learning. *BMC Genom* 22:1–10
58. Smulko J, Wróbel M (2017) Noise sources in Raman spectroscopy of biological objects. *Dyn Fluct Biomed Photon XIV* 10063:54–60 (**SPIE**)
59. Rodriguez JD, Westenberger BJ, Buhse LF, Kauffman JF (2011) Standardization of Raman spectra for transfer of spectral libraries across different instruments. *Analyst* 136(20):4232–4240
60. Chen H, Xu W, Broderick NG (2019) An adaptive and fully automated baseline correction method for Raman spectroscopy based on morphological operations and mollification. *Appl Spectrosc* 73(3):284–293
61. Klaverdijk M, Nemati M, Ottens M, Klijn ME (2025) Impact of bioreactor process parameters and yeast biomass on Raman spectra. *Biotechnol Prog* 41(6):70050
62. Dudek M, Zajac G, Szafraniec E, Wiercigroch E, Tott S, Malek K, Kaczor A, Baranska M (2019) Raman optical activity and Raman spectroscopy of carbohydrates in solution. *Spectrochim Acta Part A Mol Biomol Spectrosc* 206:597–612

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.