Methods

André Orth*, Hawo H. Höfer and Markus Reischl

Uncertainty-aware Bayesian neural networks for FTIR gas-phase process monitoring: a semi-synthetic feasibility study

Unsicherheitsquantifizierung mittels Bayes'scher Neuronaler Netze für die FTIR-gestützte Gasphasen-Prozessüberwachung: eine semi-synthetische Machbarkeitsstudie

https://doi.org/10.1515/auto-2026-0033
Received March 16, 2026; accepted October 6, 2026

Abstract: To remain safe and actionable, multi-component gas-phase Fourier-transform infrared (FTIR) process analysis requires a reliability estimate for each concentration prediction. In a semi-synthetic feasibility study for a fixed compound library, we present a Bayesian one-dimensional Convolutional Neural Network (1D-CNN) trained on spectra from single-compound reference data that predicts concentrations, compound presence probabilities and per-prediction uncertainty bounds. The model yields accurate concentration predictions (mean absolute error 0.165 in $\log_{10}$-space, within a factor of about 1.5) and uncertainty bounds that remain well-calibrated under severe instrumental artifacts. A pre-inference amplitude guard screens for inputs outside the detector's valid operating range. Validation on measured spectra is the necessary next step before deployment.

Keywords: Bayesian neural network; FTIR spectroscopy; uncertainty quantification; out-of-distribution detection; process monitoring

Zusammenfassung: Um einen sicheren und handlungsrelevanten Betrieb zu gewährleisten, erfordert die Mehrkomponenten-Gasanalyse mittels Fourier-Transform-Infrarotspektroskopie (FTIR) eine messungsspezifische Zuverlässigkeitsabschätzung zu jeder Konzentrationsvorhersage. In einer semi-synthetischen Machbarkeitsstudie für eine festgelegte Stoffbibliothek präsentieren wir ein Bayes'sches eindimensionales Convolutional Neural Network (1D-CNN), trainiert auf Spektren aus Einzelkomponenten-Referenzdaten, das Konzentrationen, Komponentenpräsenzen und messungsspezifische Unsicherheitsintervalle bestimmt. Das Modell liefert genaue Konzentrationsvorhersagen (mittlerer absoluter Fehler 0,165 im $\log_{10}$-Raum, innerhalb eines Faktors von etwa 1,5) und unter starken Messartefakten gut kalibrierte Unsicherheitsintervalle. Eine vorgeschaltete Amplitudenüberwachung schirmt Signale außerhalb des gültigen Detektormessbereichs vor der Modellauswertung ab. Die Validierung an Messspektren ist der notwendige nächste Schritt vor dem Einsatz.

Schlagwörter: Bayes'sches neuronales Netz; FTIR-Spektroskopie; Unsicherheitsquantifizierung; Ausreißererkennung; Prozessüberwachung

1 Introduction

Fourier-transform infrared (FTIR) spectrometers are widely used as process analyzers in pharmaceutical manufacturing, petrochemical refining and chemical production, where real-time gas-phase composition data are fed directly into supervisory control systems [1], [2]. In this technique, infrared light is passed through the gas sample and its absorption is measured as a function of wavenumber. The resulting absorbance spectrum contains characteristic peaks at wavenumbers corresponding to the vibrational modes of each compound present, from which their concentrations can be determined. In such automated monitoring systems, a concentration estimate is only actionable if its reliability can be assessed on a per-measurement basis. Therefore, a control system receiving an anomalous reading must be able to determine whether the deviation originates

***Corresponding author: André Orth**, Karlsruher Institute for Technology (KIT), Institute for Automation and Applied Informatics (IAI), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Baden-Württemberg, Germany, E-mail: andre.orth@kit.edu. https://orcid.org/0009-0006-1793-6026
Hawo H. Höfer and Markus Reischl, Karlsruher Institute for Technology (KIT), Institute for Automation and Applied Informatics (IAI), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Baden-Württemberg, Germany

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

from elevated measurement noise, an instrument fault or a genuine process excursion.

Partial least squares (PLS) regression remains the industrial standard for quantitative FTIR analysis [3]. PLS provides point-concentration estimates but no principled per-prediction uncertainty [4]. Consequently, automated systems must either apply conservative, manually tuned alarm thresholds that generate excessive false positives, or defer all anomaly decisions to a human operator, negating the benefit of continuous real-time monitoring [1], [2]. Deep learning methods applied to spectral data have demonstrated accuracy advantages over PLS, particularly under high noise and without preprocessing [5], [6]. Neural networks trained on synthetic spectral data have further been shown to transfer reliably to experimental measurements across multiple 1D characterization techniques, including XRD and Raman spectroscopy [7], [8] and have been applied to quantitative XPS composition analysis from synthetically generated training data [9]. Despite these advantages, the vast majority of published spectral models output scalar predictions without uncertainty quantification [10], [11]. While uncertainty-aware deep learning has been applied to other analytical measurement tasks [12], no such approach has been demonstrated for FTIR-based gas-phase monitoring. Deterministic neural networks share this limitation: while they can match or exceed PLS accuracy, they produce point predictions without calibrated per-prediction uncertainty, leaving the reliability assessment problem unsolved. An unreliable concentration estimate issued without a validity signal may either suppress an alert for a genuine hazardous process excursion or trigger unnecessary intervention on a false positive, both of which carry direct operational and safety consequences in regulated industries. The central challenge addressed in this work is therefore the absence of calibrated, per-prediction uncertainty in deployed spectral analyzers.

We present a 1D-CNN/BNN model that predicts gas-phase concentrations, compound presence and calibrated per-prediction uncertainty from FTIR absorbance spectra. The Bayesian neural network (BNN) backbone achieves this by outputting a predicted concentration value together with a confidence estimate for each compound, rather than a single value alone [13], [14]. This uncertainty separates into two components: one driven by measurement noise (aleatoric) and one reflecting how far the input lies from the training distribution (epistemic). Elevated aleatoric uncertainty indicates poor signal quality and suggests repeating or extending the measurement while elevated epistemic uncertainty flags out-of-distribution (OOD) inputs for human review before any automated action is taken. To train the model, our approach generates semi-synthetic spectra derived from certified single-compound reference spectra from the National Institute of Standards and Technology (NIST) database [15], requiring no real measurements of mixtures.

As a feasibility study for a fixed compound library, this work presents a **semi-synthetic training pipeline** combining single-compound reference spectra into synthetic mixtures with realistic instrument artifacts [7], [9]. Trained on this data, a **1D-CNN/BNN model** produces calibrated aleatoric and epistemic uncertainty estimates alongside quantitative multi-component concentration predictions from FTIR absorbance spectra. The trade-off between calibration sharpness and OOD interval behavior is analyzed through a **systematic study of uncertainty loss weighting** and three **distinct OOD failure modes** relevant to deployed process analyzers are identified and characterized. Additionally, a **deterministic amplitude guard** applied before model inference screens for signals outside the detector's valid operating range independent of the learned model.

2 Methods

2.1 Overview

As illustrated in Figure 1, our method targets real-time gas-phase composition monitoring in continuous industrial processes and separates into an offline training path and an online inference path. Superposition of single-compound reference spectra using Beer–Lambert law provides a physically grounded and computationally efficient basis for synthetic training data generation. Using this approach allows us to create large-scale mixture spectra from certified single-compound measurements without requiring real mixture samples [16] and in sufficient quantities to train our model. The training path consists of two stages. The first stage is spectral data generation, where NIST reference spectra [15] are combined via Beer–Lambert superposition into synthetic mixture spectra, which are then superimposed with instrument-realistic artifacts in three severity and one out-of-distribution (OOD) tier. The second stage is model training, in which a one-dimensional convolutional neural network (1D-CNN) with Squeeze-and-Excitation (SE) channel attention encodes each spectrum into a compact descriptor fed to a Bayesian neural network (BNN) backbone. The BNN backbone produces per-compound concentration estimates together with per-prediction uncertainty bounds, distinguishing irreducible measurement noise (aleatoric) from model-confidence uncertainty arising from out-of-distribution inputs (epistemic) and per-compound presence probabilities. The online inference path applies the

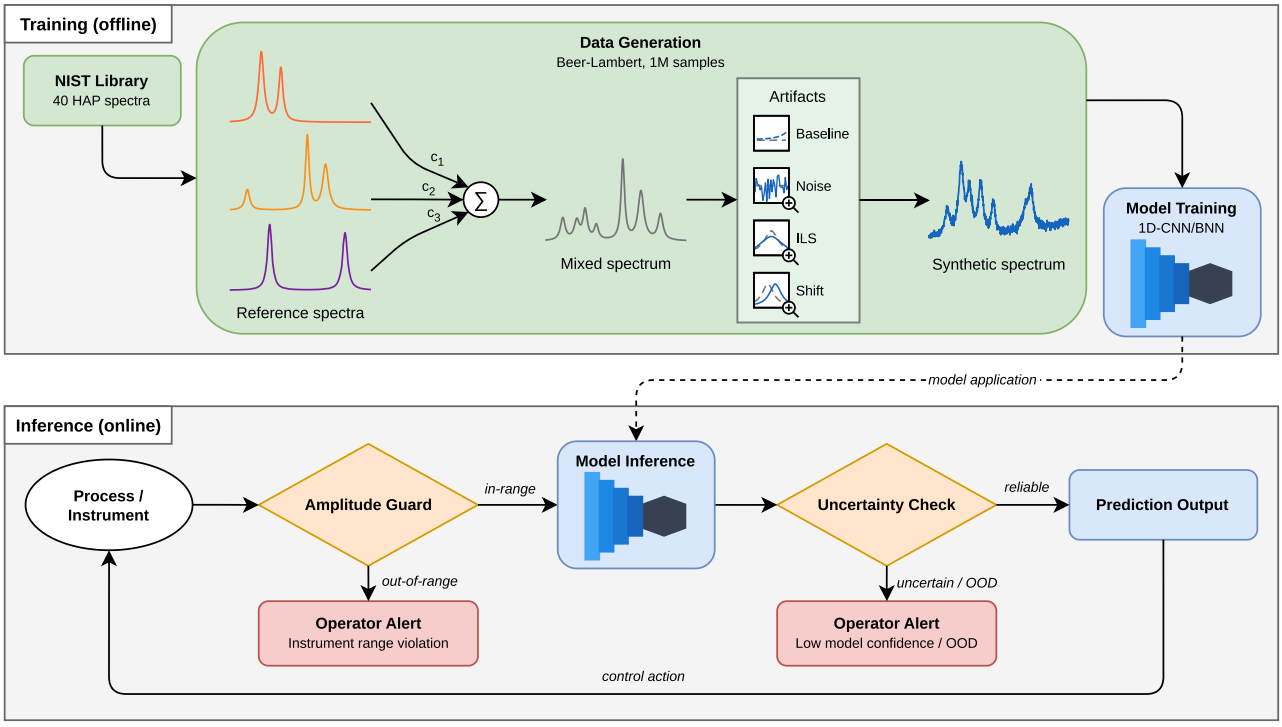


Figure 1: Overview of the proposed pipeline. Top: **offline training** path with a data generation stage based on NIST single-compound reference spectra combined via Beer–Lambert superposition and subsequently augmented with instrument-realistic artifacts (baseline drift, detector noise, instrument line shape (ILS) broadening and wavenumber shift) to create final synthetic mixture spectra. The synthetic data is then used to train the 1D-CNN/BNN model. Bottom: **online inference** path with deterministic amplitude guard, calibrated 1D-CNN/BNN inference and uncertainty-based operator alertings.

trained 1D-CNN/BNN to incoming spectra at runtime, preceded by a deterministic amplitude guard that screens for signals outside the detector's valid operating range and returns an operator alert. After model prediction, an uncertainty check flags out-of-distribution inputs for operator review before any control action is sent back to the supervised process. Calibrated uncertainty output is essential in this context as per-measurement reliability assessment is a core requirement of automated process control, motivating both the BNN backbone and the amplitude guard described in the following sections.

2.2 Semi-synthetic dataset generation

Our data generation pipeline proceeds in five steps: reference spectrum preprocessing, mixture generation, training label assignment, instrument artifact simulation and dataset structure.

2.2.1 Reference spectra and preprocessing

The NIST Quantitative Infrared Database [15] was used as a source to obtain single-compound absorption spectra. The database provides experimentally measured absorbance spectra of 40 US EPA Hazardous Air Pollutants (HAPs), recorded at 296 K and 1 atm. The 3-term Blackman-Harris apodization at 1.0 cm^{-1} nominal resolution was selected to match the typical specification of industrial process FTIR analyzers [16], ensuring that the synthetic training data is representative of real deployment conditions. The NIST spectra are provided at differing native wavenumber ranges. All 40 reference spectra were interpolated onto a common wavenumber range of $\tilde{\nu}$ = 600–4,000 cm^{-1} at $\Delta\tilde{\nu}$ = 0.964 cm^{-1}, resulting in 3,528 data points per spectrum. The NIST values, given per ppm and meter path length, were converted to molar absorption coefficients (L mol^{-1} cm^{-1}) via the ideal gas law (298 K, 1 atm).[1] The model operates on absorbance spectra delivered by the analyzer after the standard single-beam background ratio, so no additional background correction is applied at the model level. Residual drift from infrequent background updates is captured by the baseline artifact tier.

1 Ppm by volume is the standard concentration unit in industrial process gas monitoring and regulatory emission reporting [1]; molar and ppm concentrations differ by a constant factor, so all log-space metrics are identical in both units.

2.2.2 Mixture generation

For each sample, the number of active compounds $K \in [1, 10]$ is drawn from a discrete probability distribution reflecting realistic mixture complexity. It is set up so that single-component spectra account for 20 % of samples to support compound identification and two-to-four component mixtures for 55 % as the most common process scenario. The remaining 25 % consists of five-component mixtures (10 %) and six-to-ten component stress tests (15 %). Concentrations c_i for each active compound are sampled log-uniformly over three decades (the upper bound of each compound is set so that its strongest band reaches 0.80 AU at 10 cm path length, the lower bound is 10^{-3} of it. Resulting lower bounds 0.2–100 ppm, upper bounds 220 ppm to 10 %). Log-uniform was chosen because linear-uniform sampling would allocate the majority of samples to high-concentration events, severely undersampling the low-concentration regime where detection is most critical. Spectra are then assembled by Beer–Lambert superposition, the physical basis of quantitative FTIR analysis in dilute gas-phase mixtures under the assumption of linear detector response [1], [16] (see Appendix A for the full formulation and unit conversion).

2.2.3 Training labels

Each training sample encodes K active and $40 - K$ absent compounds, where K is the number of compounds drawn for that sample from the mixture-complexity distribution defined above and $40 - K$ is the remainder of the fixed 40-compound hazardous air pollutant (HAP) library. Active compounds receive a $\log_{10}$-space concentration label. Absent compounds are assigned a sentinel value of -99 and a binary presence label of zero. The sentinel is chosen in $\log_{10}$-space to lie far below any physically valid concentration, so the loss mask condition $y_{ij} > -50$ cleanly separates present from absent compounds without risk of collision with real low-concentration labels. These labels are stored per-sample in the dataset and are used directly by the loss masking described in Section 2.3.

Because absorbance accumulates additively across all K active compounds, the peak of $A(\tilde{\nu})$ can exceed the detector's quantitative range even when each individual c_i is within its sampled bounds. Simple amplitude normalization is not applicable here as it would destroy the absolute concentration information encoded in the spectral amplitudes via Beer–Lambert law. Instead, spectra with $A_{\max} > 1.0$ AU and their concentration labels were scaled simultaneously by $0.95/A_{\max}$, simulating a physical dilution that preserves the Beer–Lambert self-consistency of each training pair while avoiding detector nonlinearity [16]. The pre-dilution peak amplitude $A_{\max,\mathrm{pre}}$, the noise-free Beer–Lambert peak before dilution and before any artifact, was stored per sample for use by the amplitude guard described in Section 2.5. The virtual dilution is part of training-data generation only, represents a physical sample dilution and is not applied at inference.

2.2.4 Instrument artifact simulation

As shown in Figure 1, four artifact classes were superimposed following literature-grounded parameters for industrial process FTIR analyzers [15], [16] using an instrument-characterization-based simulation approach [17], [18] to improve model generalization to spectra acquired under non-ideal measurement conditions.

- Additive linear, quadratic or exponential baseline drift, partly with a localized band at 3,275 cm^{-1},
- white Gaussian detector noise parameterized by the signal-to-noise ratio (SNR),
- additional Gaussian instrument line shape (ILS) broadening beyond the resolution of the reference spectra and
- wavenumber axis shift by interpolation, combined with an intensity scaling of up to ±5 %.

These four artifact classes are applied at three in-distribution and one out-of-distribution severity tiers, covering a wide range of instrumentation conditions from controlled laboratory measurements to severely degraded field analyzers. Artifact parameters are summarized in Table 1 and example spectra are shown in Figure 2. The clean tier represents controlled measurement conditions with well-maintained instrumentation, the moderate tier corresponds to typical industrial acquisition practice and the severe tier reflects degraded measurement quality approaching operational limits. The OOD-extreme tier lies entirely outside the training distribution and is used exclusively for held-out OOD evaluation. The in-distribution tiers are sampled with probabilities of 20 %, 50 % and 30 % for clean, moderate and severe, respectively, deliberately weighting toward degraded conditions to prevent the model from overfitting to noise-free spectra.

2.2.5 Dataset structure

The total dataset was exported with one million samples split into training, validation and test sets using a ratio of 80/10/10. The three in-distribution tiers are sampled at 20 %/50 %/30 % across all splits. Three additional held-out OOD test sets were generated separately: `ood_extreme` (300 spectra of in-distribution compounds rendered with OOD-extreme artifacts), `ood_conc` (200 spectra with concentrations a factor of 10^2 to 10^4 below or above the training bounds) and `ood_chemical` (300 spectra of five compounds withheld

Table 1: Artifact severity tiers used for training, in-distribution testing and OOD evaluation. SNR ranges are given in dB, wavenumber shift in cm^{-1}. Drift coefficients are uniform in $\pm 0.02 \cdot c_s$ AU. Drift is applied with probability 0.5 at the moderate tier and always at the severe and OOD-extreme tiers. ILS broadening is applied with tier-dependent probability (0.1/0.2/0.8/1.0).

Tier	SNR /dB	Drift scale c_s	Extra ILS /cm^{-1}	Wn. shift /cm^{-1}
Clean	60–80	None	0–0.3	0
Moderate	40–60	1	0–0.5	±0.10
Severe	20–40	2–4	0.3–1.0	±1.0
OOD-extreme	10–20	4	1.5–2.0	±4.0

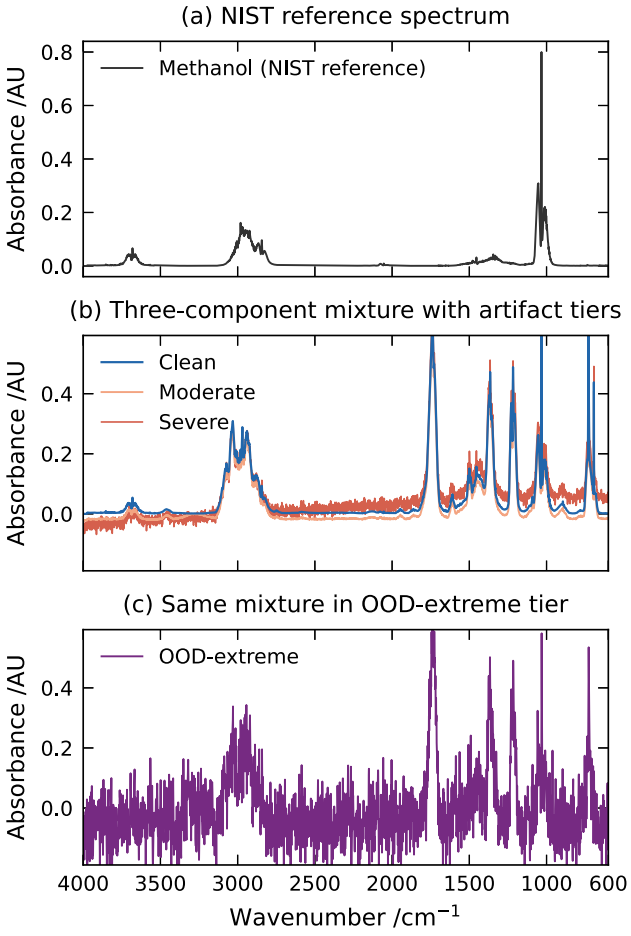


Figure 2: Semi-synthetic FTIR spectra illustrating the data generation pipeline. (a) Single-compound methanol reference spectrum from the NIST database [15] before superposition. (b) Example multi-component mixture spectrum at the clean (blue), moderate (orange) and severe (red) artifact tiers. Progressive broadening and baseline curvature increase with tier severity. (c) The same mixture at the OOD-extreme tier, outside the training distribution and used exclusively for held-out OOD evaluation.

from training: CCl_4, SO_2, *n*-butane, ethyl acrylate and benzene). The remaining 35 compounds are used for training and in-domain evaluation. These five compounds were selected to span distinct chemical families (halogenated, sulfurous, aliphatic, acrylate, aromatic) while sharing absorption band regions with the training set, providing a stringent test of closed-set generalization under chemical novelty. To stabilize and speed up training, all spectra were z-score normalized using the training-set mean and standard deviation prior to model input.

2.3 Model architecture and training

Our model approach produces three outputs per input spectrum: **per-compound concentration estimates**, **aleatoric uncertainty bounds** and **per-compound presence probabilities**. Each output type has dedicated labels, a loss term and evaluation metrics, defined in the following. A simplified display of the architecture is shown in Figure 3.

2.3.1 Architecture

The 3528-channel absorbance spectrum is encoded by four convolutional blocks with filter sizes of 32, 64, 128 and 256 and a kernel width of $k = 9$ ($\approx 8.7\ cm^{-1}$), chosen to span the full peak shape of gas-phase absorption bands at $1\ cm^{-1}$ resolution [5], [19]. After each pooling step, a Squeeze-and-Excitation (SE) block with reduction ratio $r = 4$ re-weights filter channels on a per-sample basis to suppress spectral sub-bands corrupted by baseline drift or ILS broadening [20]. A GlobalAveragePooling1D operation reduces the final block output to a 256-dimensional spectral descriptor.

The 256-dimensional descriptor is passed through two fully-connected layers (256 and 128 units). Monte Carlo (MC) Dropout [13] with rate 0.2, applied unconditionally during both training and inference via a custom `AlwaysDropout` layer,[2] is placed after each convolutional block and after each fully-connected layer. MC Dropout was selected for competitive calibration at lower inference overhead, relevant for real-time deployment constraints. At test time, $T = 50$ independent stochastic forward passes are performed per input spectrum across three independently trained models. The aleatoric variance is estimated as the mean predicted variance across T passes and the epistemic variance as the inter-pass variance of predicted means [14], [21]. Only $\hat{\sigma}_{\text{alea}}$ is scaled by the conformal factor. $\hat{\sigma}_{\text{epi}}$ is reported separately as

2 This custom layer overrides the Keras default behavior of disabling dropout at inference time.

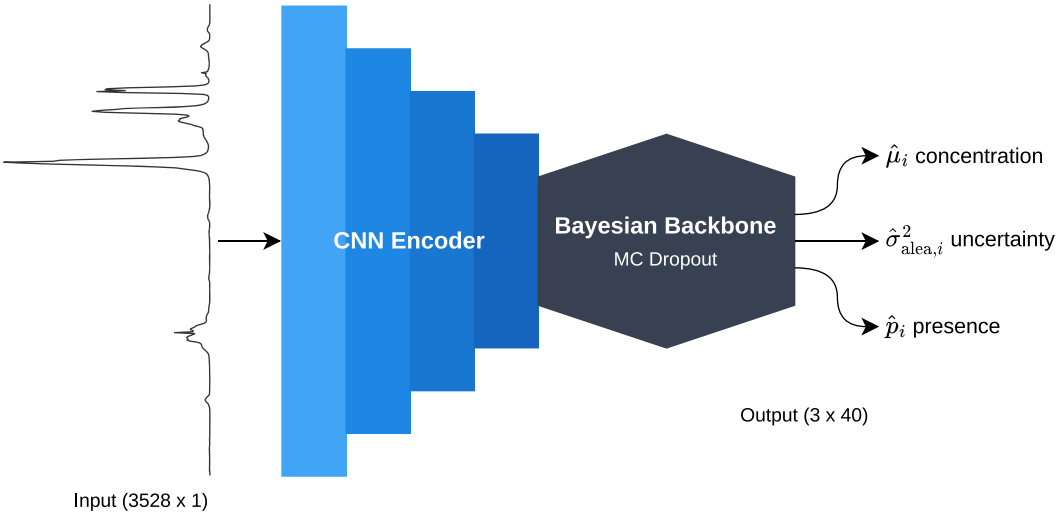


Figure 3: Architecture of the 1D-CNN/BNN model. The 3528-channel absorbance spectrum is encoded by four convolutional blocks (blue). A BNN backbone with Monte Carlo (MC) dropout (dark grey) produces $T = 50$ independent stochastic forward passes per inference call. Three output heads yield per-compound concentration means $\hat{\mu}_i$ and log-variances $\hat{\sigma}^2_{\text{alea},i}$ in $\log_{10}$ (ppm) and binary presence probabilities $\hat{p}_i$.

a model-confidence signal. Full per-layer specifications are provided in Table 7.

A scalar noise amplitude estimate $\overline{\delta}$ is computed from the input spectrum and provided as an additional input to the log-variance head, supplying an explicit noise signal that ties aleatoric uncertainty to the measured noise level. It is defined as the mean absolute first difference of the spectrum,

$$\overline{\delta} = \frac{1}{N-1} \sum_{n=1}^{N-1} |A_{n+1} - A_n|, \tag{1}$$

which scales directly with additive noise amplitude. Three output heads produce per-compound concentration means $\hat{\mu}_i$, log-variance values $\hat{\sigma}^2_{\text{alea},i}$ (in log space, conditioned on $\overline{\delta}$) and presence probabilities $\hat{p}_i$ respectively (full specifications in Table 7). This architecture is referred to as *CNN-BNN-SE*. The model class as a whole, encompassing variants with and without SE attention, is referred to as the *1D-CNN/BNN* model. The ablation baseline without SE blocks is referred to as *CNN-BNN*.

2.3.2 Training

Because the concentration and presence heads operate on fundamentally different output spaces, namely heteroscedastic regression over continuous log-space concentrations and weighted binary classification over presence probabilities, a composite loss is required with one dedicated term per output type. For concentration regression, the standard heteroscedastic negative log-likelihood (NLL) loss penalizes both prediction error and uncertainty miscalibration, but is prone to log-variance collapse, where the model reduces loss by inflating $\hat{\sigma}_{\text{alea}}$ rather than improving predictions. To prevent this, a scalar hyperparameter $\beta \in [0, 1]$ is introduced into the variance weighting term $\hat{\sigma}^{-2\beta}$ of the NLL. At $\beta = 0$ the loss reduces to mean squared error and the variance head receives no gradient while at $\beta = 1$ it recovers the standard NLL. For intermediate values, a stop-gradient is applied to $\hat{\sigma}^{-2\beta}$, decoupling prediction-accuracy gradients from variance-calibration gradients. This is referred to as the β-**NLL** loss [22]. The choice of β has safety-relevant consequences for OOD interval behavior. The masked regression loss is computed only over present compound-sample pairs:[3]

$$\mathscr{L}_{\text{reg}} = \frac{1}{|M|} \sum_{(i,j)\in M} \left[\frac{(y_{ij} - \hat{\mu}_{ij})^2}{2\hat{\sigma}^2_{ij}} + \frac{1}{2} \log \hat{\sigma}^2_{ij} \right] \cdot \text{sg}\left(\hat{\sigma}^{2\beta}_{ij}\right), \tag{2}$$

with sg(·) denoting stop-gradient and $\beta = 0.50$. The presence term $\mathscr{L}_{\text{pres}}$ is weighted binary cross-entropy,[4] giving the total loss $\mathscr{L} = \mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{pres}}$. To assess sensitivity to β, models were additionally trained with $\beta \in \{0.25, 0.75\}$ with all other settings held fixed. Training hyperparameters are listed in Table 7.

3 Absent compounds carry a $\log_{10}$-space sentinel label of −99 and are excluded by the mask condition $y_{ij} > -50$, where −50 is safely above the sentinel but below any physically valid concentration.

4 The approximately 9:1 absence-to-presence class imbalance is addressed by a positive class weight of 6.0.

2.3.3 Design targets and metrics

Unless stated otherwise, all concentration metrics are reported in $\log_{10}$ (ppm).

2.3.3.1 Concentration

The regression head (a task-specific output sub-network that maps the shared encoder representation to a scalar prediction) produces concentration estimates alongside per-prediction uncertainty bounds. Concentration errors in process monitoring are multiplicative. A factor-of-two deviation at 10 ppm carries the same operational weight as one at 1,000 ppm and $\log_{10}$-space metrics capture this naturally. Regression performance is quantified by **Mean Absolute Error (MAE)** as the primary metric, **coefficient of determination** (R^2) as a scale-free variance-explained score and **Root Mean Square Error of Prediction (RMSEP)** as a secondary reported value. Design targets are MAE ≤ 0.30 (corresponding to factor-of-two concentration accuracy, sufficient for binary alarm gating where the decision threshold is well above measurement uncertainty) and $R^2 \geq 0.95$.

2.3.3.2 Aleatoric uncertainty

The uncertainty head produces a predicted uncertainty interval alongside each concentration estimate. After rescaling the intervals to match the nominal coverage level (see Section 2.4), the interval quality is assessed by the **Prediction Interval Coverage Probability (PICP$_a$)**, defined as the fraction of true concentration values falling within the reported bounds. A value of PICP$_a$ = 0.95 means 95 % of true values are covered, as required by continuous emission monitoring standards [23]. The **total prediction interval coverage probability** (PICP) additionally incorporates the unscaled epistemic term $\widehat{\sigma}^2_{\text{epi}}$ and serves as an upper-bound safety diagnostic. Whether the model additionally tracks measurement noise, i.e. produces wider intervals for noisier spectra rather than a fixed learned width, is quantified by the **aleatoric responsiveness** ρ_{alea}: the Pearson correlation between $\widehat{\sigma}_{\text{alea}}$ and the injected noise standard deviation σ_n (target: $\rho_{\text{alea}} \geq 0.55$). This diagnostic is only feasible in the semi-synthetic domain where σ_n is known. The noise amplitude estimate $\overline{\delta}$, introduced above, provides the explicit coupling between $\widehat{\sigma}_{\text{alea}}$ and the measured noise level.

2.3.3.3 Compound presence

The presence head produces a probability $\widehat{p}_i \in [0, 1]$ for each compound, indicating whether it is present in the input spectrum. A compound is labeled present ($p_i = 1$) if its $\log_{10}$-space label exceeds the sentinel threshold of −50 and absent ($p_i = 0$) otherwise. Discrimination performance is evaluated by the **Area Under the Receiver Operating Characteristic Curve (AUROC)** [24], computed over all pooled compound-sample pairs. AUROC is threshold-independent and therefore suitable for architecture comparison. The target AUROC ≥ 0.90 is set to ensure presence/absence gating within the fixed library of 35 training compounds.

These three design targets frame the evaluation in Section 3 and are summarized in Table 2.

2.4 Uncertainty calibration

Split conformal prediction [25], a distribution-free post-hoc interval scaling method, is applied to the validation set. For each present compound-sample pair, the prediction error is normalized by the predicted aleatoric uncertainty, measuring how many uncertainty units separate prediction from truth. The scale factor $\widehat{q}$ is then set to the empirical 95th percentile of these normalized errors across all n_{cal} validation pairs. At inference, prediction intervals are formed as $\widehat{\mu}_i \pm \widehat{q}\,\widehat{\sigma}_{\text{alea},i}$, meaning the learned uncertainty is scaled by the factor needed to reach the nominal coverage level. This yields a finite-sample lower bound of at least 95 % marginal coverage, averaged over all present compound-sample pairs, without retraining or distributional assumptions, provided calibration and test pairs are exchangeable. The bound is marginal: it does not give coverage per compound, concentration range or artifact tier and it does not hold for distribution-shifted inputs such as the OOD splits or spectra from a real instrument, for which coverage is reported empirically only. Full technical details, including the finite-sample correction and exchangeability assumption, are given in Appendix C.

The term *calibration* is used here in the statistical sense: adjusting predictive intervals post-hoc so that their empirical coverage matches the nominal level, distinct from instrument calibration against a reference standard or a Guide to the Expression of Uncertainty in Measurement (GUM) style uncertainty budget, which is beyond the scope of this work.

Table 2: Primary design targets for the CNN-BNN-SE model.

Output	Metric	Target
Concentration	MAE	≤ 0.30
	R^2	≥ 0.95
Aleatoric coverage	PICP$_a$	≥ 0.95
Noise tracking	ρ_{alea}	≥ 0.55
Presence detection	AUROC	≥ 0.90

2.5 Deployment safety and OOD evaluation

Before neural network inference, a deterministic **amplitude guard** screens each incoming spectrum for signals outside the detector's valid operating range. It checks the pre-dilution peak absorbance $A_{\mathrm{max,pre}}$ against two fixed thresholds: inputs below 10^{-5} AU are flagged `below_range` (signal absent or detector fault) and inputs above 10.0 AU are flagged `above_range` (extreme absorbance or instrument fault). Because no dilution is applied at inference, $A_{\mathrm{max,pre}}$ corresponds to the peak absorbance of the incoming measured spectrum and requires no information beyond the spectrum itself. In the semi-synthetic evaluation, the guard is applied to the noise-free Beer–Lambert peak before dilution and artifacts, an idealization whose consequences for both thresholds are discussed below. Flagged spectra are returned to the process control system without model evaluation, preventing out-of-range inputs from silently producing faulty concentration estimates.

The upper threshold is intentionally set well above the detector's quantitative range to serve as a coarse screen for spectra with anomalously high absorbance, whether caused by extreme sample concentrations exceeding any physically meaningful measurement range or by instrument faults such as a stuck analog-to-digital converter, a disconnected measurement cell or a fully blocked optical path. The instrument-specific quantitative range, typically below approximately 1.0 AU and increasingly nonlinear toward 2.0–3.0 AU due to stray light and zero-offset effects, is not enforced by a hard threshold because it varies across instruments and a fixed cutoff near this range would introduce false rejections on instruments with wider quantitative response. On measured spectra, absorbance saturates at a few AU due to stray light, so the 10.0 AU default flags only invalid readings. Over-concentration on a given instrument can therefore only be flagged by setting the upper threshold to that instrument's photometric ceiling at commissioning. On the semi-synthetic data, recall and false-positive rate are unchanged for any upper threshold between 1.9 and 62 AU. The lower threshold does not transfer directly. Applied to the simulated artifact-corrupted spectrum, the same check would flag 2.25 % of the 100,000 in-domain test spectra (negative baseline drift pushes the spectrum maximum below 10^{-5} AU) but only 8 of the 32 below-range `ood_conc` spectra. A deployed lower check therefore has to compare the peak height above a baseline estimate with the noise level, for example $\overline{\delta}$ (Equation (1)), a variant not evaluated here. Spectra between the thresholds but outside the training-domain amplitude (noise-free training peaks do not exceed 1.0 AU, consistent with reported quantitative ceilings of approximately 0.8–1.0 AU for process FTIR) are not flagged by the guard. The model's uncertainty may respond to such inputs, but this regime was not evaluated.

To characterize model behavior under distribution shift, the metrics defined in Section 2.3 are evaluated across all three OOD sets alongside one additional signal. It quantifies how far each spectrum's internal encoder representation deviates from the cloud of in-distribution spectra through a feature-space distance score. Spectra from unknown compounds or instrument conditions outside the training range are expected to map to remote regions of this space, producing elevated scores [26]. Full implementation details are given in Appendix D.

3 Results and discussion

3.1 Overview

The proposed gas-phase analyzer model (CNN-BNN-SE) produces three outputs per input spectrum: a per-compound concentration estimate, a per-prediction uncertainty bound for alarm and actuation gating and a per-compound presence probability for closed-set species identification. Sections 3.2 and 3.3 characterize nominal in-domain performance, Section 3.4 motivates the key architectural choices, Section 3.5 evaluates robustness under distribution shift and Section 3.6 defines the scope and limitations of these results.

3.2 In-domain accuracy and compound presence detection

The MAE and presence-detection targets are met, while R^2 (0.931) remains below its target of 0.95, as summarized in Table 2. The complete in-domain metrics are listed in Table 3.

The concentration accuracy target is met with substantial margin. In practical terms, a species truly present at 100 ppm is typically reported between 68 and 146 ppm, well within the factor-of-two threshold required for threshold-based alarming. R^2 weighs squared errors and is dominated by a minority of large errors: the largest 5 % of absolute errors contribute 49 % of the squared error sum. These errors occur mainly at the lower end of each compound's concentration range, where the model over-predicts as the signal approaches the noise level. Accuracy degrades gracefully from clean to severe artifact conditions with no tier-dependent offset, indicating that performance loss under degraded instrumentation is proportional to measurement quality. Compound presence detection exceeds its

Table 3: In-domain accuracy and presence-detection metrics for the CNN-BNN-SE model. MAE, RMSEP and R^2 are reported for all spectra together and tier-stratified MAE rows show how prediction errors increase from clean to severe artifact conditions. Presence AUROC quantifies compound-wise detection performance.

Metric	Value	Target
MAE	0.165	≤0.30
RMSEP	0.267	–
R^2	0.931	≥0.95
MAE, clean	0.126	–
MAE, moderate	0.143	–
MAE, severe	0.228	–
Presence AUROC	0.923	≥0.90

AUROC target, providing closed-set species gating across the compound library.

3.3 Calibration performance

The raw model uncertainty output requires post-hoc recalibration to reach the nominal 95 % coverage level. This is achieved via split conformal prediction [25], which provides a finite-sample lower bound of at least 95 % marginal coverage without distributional assumptions and without retraining, provided calibration and test pairs are exchangeable (see Appendix C for full details). Calibration results are shown in Figure 4.

On the held-out test set, coverage is 0.946 (Table 6), 0.4 percentage points below the nominal level and the PICP_a target of 0.95 (validation coverage is 0.950 by construction). Conditional coverage is lower in the upper part of each compound's concentration range (0.859), which is where alarm thresholds typically lie. Conditional coverage details and the approximate exchangeability of calibration and test data are discussed in Appendix C. The uncertainty output actively widens for noisier input spectra rather than producing a fixed learned width (ρ_{alea}), so that each reported bound responds to the measurement quality of the individual spectrum. This per-measurement reliability is a core requirement for automated process control. Beyond the nominal aleatoric coverage, total coverage incorporating the epistemic term (0.999 on the in-domain test set) provides an additional conservative margin for in-domain inputs. It does not prevent overconfident predictions for chemically novel inputs (Section 3.5). The two uncertainty components are reported separately, allowing the control system to distinguish between measurement-driven and model-confidence-driven reliability signals independently.

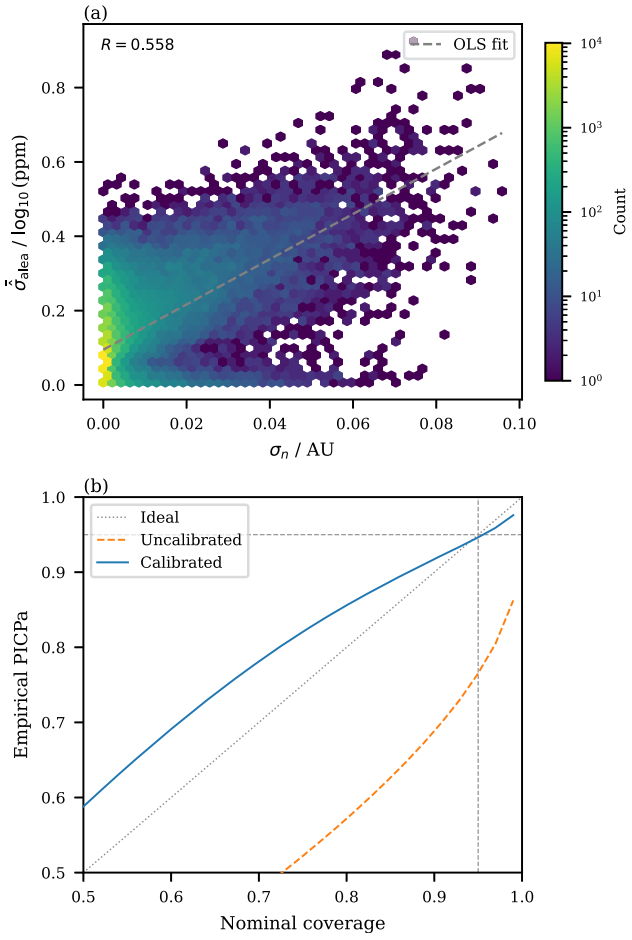


Figure 4: Uncertainty calibration results on the in-domain test set. (a) Predicted aleatoric standard deviation $\bar{\sigma}_{\mathrm{alea}}$ versus injected noise standard deviation σ_n across 100,000 test samples ($\rho_{\mathrm{alea}} = 0.558$, $p < 0.001$). A positive correlation confirms that the model elevates uncertainty for noisier inputs rather than producing a learned constant. (b) Reliability diagram comparing calibrated (post-conformal) and uncalibrated coverage as a function of nominal level. The calibrated curve meets the diagonal at the 95 % level used for calibration and lies above it at lower nominal levels, while the uncalibrated output undercovers at all levels.

3.4 Ablation studies

The two architectures defined in Section 2.3, CNN-BNN-SE and CNN-BNN, are compared under identical training conditions. Full results are listed in Table 4.

In-domain concentration accuracy is comparable between both variants. The main difference is prediction-interval validity (PICP_a) under severely degraded instrumentation. CNN-BNN-SE maintains calibrated coverage under `ood_extreme` conditions while CNN-BNN falls below the 95 % nominal target with a gap of 5.5 percentage points. This improvement is attributed to per-sample spectral channel reweighting by the SE blocks [20], which adaptively suppresses sub-bands corrupted by baseline drift or ILS broadening, a capability the

Table 4: Architecture ablation on the 100,000-sample test set and `ood_extreme` split (β = 0.50). PICP$_a$ (ood) quantifies prediction-interval validity under severely degraded conditions. ρ_{alea} is the Pearson correlation between $\widehat{\sigma}_{\text{alea}}$ and injected noise on the in-domain test set. Bold values indicate the better architecture per metric (meeting the 0.95 nominal level for coverage metrics).

Architecture	MAE	R^2	PICP (ood)	PICP$_a$ (ood)	ρ_{alea}
CNN-BNN	0.176	0.924	0.935	0.895	**0.596**
CNN-BNN-SE	**0.165**	**0.931**	**0.986**	**0.958**	0.558

baseline architecture cannot provide. CNN-BNN-SE is therefore selected as the primary architecture, accepting a marginal reduction in noise-tracking responsiveness (ρ_{alea}) as a justified trade-off against interval coverage under OOD conditions.

The choice of β has direct safety consequences for deployment. At β = 0.25, the conformal step shrinks all intervals ($T_{\text{equiv}} < 1$). On `ood_extreme` they widen less than the prediction errors grow, so coverage drops to 0.502, rendering the model unsuitable for deployment. At β = 0.75, the raw standard deviation underestimates the errors by about an order of magnitude and the calibrated intervals are wider than at β = 0.50 (mean in-domain width 1.85 vs 1.37 in $\log_{10}$ units) without higher OOD coverage. At β = 0.50, PICP$_a$ under OOD conditions is highest among the three evaluated settings (0.895 vs 0.502 and 0.872) without the strong interval inflation at β = 0.75 (T_{equiv} 9.92 vs 2.01), and β = 0.50 is therefore selected among the evaluated values. The conformal scale factor T_{equiv} (see Appendix C) provides the quantitative basis for this safety assessment and is listed in Table 5.

3.5 OOD robustness and amplitude guard

Table 6 reports model performance across the in-domain test set and the three held-out OOD splits. Each split represents a qualitatively distinct failure scenario with different implications for safe deployment.

Under severe instrumental degradation (`ood_extreme`), PICP$_a$ remains at 0.958 while aleatoric uncertainty rises by a factor of 1.56, consistent with the SE ablation in Section 3.4. The feature-space distance score does not separate `ood_extreme` from in-domain spectra on average; only spectra with noise amplitudes beyond the training range are mapped far outside the in-domain feature distribution (Appendix D).

Table 5: β-NLL ablation on the CNN-BNN architecture. T_{equiv} is the conformal scale factor (see Appendix C). Values below 1 mean that the conformal step shrinks the raw intervals, including those of OOD inputs. PICP$_a$ (ood) is evaluated on `ood_extreme`. ρ_{alea} is the Pearson correlation with injected noise on the in-domain test set. For β = 0.75, the scale factor is the empirical 95th percentile of the normalized validation residuals (difference to the conformal quantile below 0.1 %).

β	T_{equiv}	MAE	PICP$_a$ (ood)	ρ_{alea}
0.25	0.77	0.175	0.502	0.533
0.50	2.01	0.176	0.895	0.596
0.75	9.92	0.176	0.872	0.782

Table 6: OOD evaluation of the SE primary model (in-domain: 100,000 spectra, OOD splits: 200–300 spectra each). The aleatoric ratio is normalized to the in-domain test mean. Values above 1 indicate active uncertainty elevation. Dist. AUROC is the OOD detection performance of the feature-space distance score in the 256-dimensional Global-AveragePooling1D (GAP) feature space (Appendix D). Pres. AUROC quantifies closed-set compound-presence detection. `ood_conc` metrics reflect unguarded inference. The amplitude guard flags all 200 spectra of an independently generated `ood_conc` set before model evaluation.

Split	MAE	PICP$_a$	$\overline{\sigma}_{\text{alea}}$	Dist. AUROC	Pres. AUROC
In-domain	0.165	0.946	1.00	–	0.923
`ood_extreme`	0.740	0.958	1.56	0.535	0.612
`ood_chemical`	5.309	0.098	1.07	0.246	0.031
`ood_conc`	3.362	0.131	1.18	0.512	0.776

For chemically novel inputs (`ood_chemical`), all available OOD signals fail simultaneously (distance-score AUROC 0.246). The model treats withheld compounds as in-domain because they share absorption band regions with the training set, making them indistinguishable at the feature level. Chemical novelty is therefore not detectable through output- or feature-space statistics alone and the model is applicable only within a fixed, known compound library [26]. Concretely, an interferent compound absent from the training library will not be flagged by any of the available OOD signals and its spectral contribution will be silently absorbed into the concentration estimates of spectrally overlapping library compounds. Open-set compound identification is identified as a necessary extension before deployment in processes where the full chemical inventory cannot be guaranteed.[5]

Without the guard, concentrations far outside the training range produce large errors (MAE 3.362) that the intervals cover in only 13 % of cases (Table 6), so such inputs have to be intercepted before inference. In the semi-synthetic evaluation, the amplitude guard flags all 200 `ood_conc` spectra of an independently generated set before inference (32 below range, 168 above range), with no false positives among 100,000 in-domain spectra (95 % upper confidence bound 3.0×10^{-5}, rule of three). The result refers to the idealized noise-free peak (Section 2.5) and moderate extrapolation inside the thresholds is not flagged.

5 Open-set classification approaches for spectral data exist. However, the present work addresses calibrated uncertainty quantification and OOD detection for a known, fixed compound library.

3.6 Scope and deployment limitations

The results reported throughout this section are bounded to the semi-synthetic evaluation domain. The NIST absorption coefficients used as the spectral basis are metrologically traceable reference spectra measurements [15], grounding the spectral physics in experiment; the predicted concentrations inherit this physical grounding but are not themselves metrologically traceable, as the learned model introduces errors for which no formal uncertainty budget has been established. Only multi-component mixture combinations and instrumental artifact parameters are modeled. The artifact tiers cover the performance range of standard to severely degraded gas-phase FTIR instrumentation, but several instrument-specific systematic effects are not yet represented in the artifact model [16]: cell window fouling, proprietary apodization functions, gas-pressure variation, gas-temperature variation and instrument drift, which affect absorption band intensities, shapes and baseline stability. Pressure- and temperature-dependent line shape and intensity changes are physically well-characterized and represent a natural extension of the synthetic data generation pipeline as additional artifact tiers, following the same approach used for the existing artifact classes. Instrument drift, cell window fouling and proprietary apodization functions are inherently instrument- and time-specific and are better addressed through transfer calibration on a small set of real spectra from the deployed instrument, rather than through synthetic simulation.

Real-instrument validation on a specific analyzer, together with transfer calibration from the NIST-trained model, constitutes the necessary next step before industrial deployment. Transfer of the NIST-trained model to a real instrument and between instruments of the same make and model, has not been demonstrated and is expected to require calibration to account for instrument-specific physical variability.

4 Conclusions

This work presented CNN-BNN-SE, a 1D-CNN/BNN model with Squeeze-and-Excitation attention, trained on NIST-grounded semi-synthetic FTIR spectra for multi-component gas-phase concentration prediction with calibrated uncertainty quantification, as a semi-synthetic feasibility study for a fixed compound library. The MAE, ρ_{alea} and presence-detection targets are met, while R^2 (0.931) remains below its target of 0.95 and aleatoric test coverage is 0.946 against the nominal 0.95. The model additionally provides calibrated per-prediction uncertainty bounds and compound presence detection. These capabilities are absent from established PLS-FTIR methods [3], [4] and position CNN-BNN-SE as a candidate for uncertainty-aware monitoring within a fixed compound library, pending validation on measured spectra.

SE channel attention contributes to prediction-interval validity under severe instrumental degradation: per-sample spectral sub-band reweighting in CNN-BNN-SE preserves interval coverage on `ood_extreme` (0.958) where CNN-BNN falls short of the nominal 95 % coverage target (0.895). A systematic β-NLL ablation establishes a safety-relevant design constraint. β = 0.25 yields $T_{\text{equiv}} < 1$, so that the calibration shrinks the intervals and coverage under severe artifacts drops to 0.502, while β = 0.75 yields wider intervals without higher coverage than β = 0.50. A deterministic pre-inference amplitude guard flags all 200 spectra of an independently generated `ood_conc` set before inference, with no false positives among 100,000 in-domain spectra, evaluated on the idealized noise-free peak. On measured spectra, both thresholds have to be referenced to the specific instrument (Section 2.5).

Chemical novelty detection remains an open problem, as none of the evaluated signals, including the feature-space distance score, detects withheld compounds. Extensions toward open-set compound identification are a natural next step. Once validated on measured spectra, the calibrated per-prediction uncertainty and OOD flagging capability would be relevant in safety-critical applications such as continuous emission monitoring in chemical production, regulatory threshold surveillance for hazardous air pollutants and automated alarm gating in pharmaceutical manufacturing, where the ability to distinguish process excursions from instrument faults or poor signal quality on a per-measurement basis is operationally decisive. Real-instrument validation and transfer calibration to a specific process analyzer, addressing the currently undemonstrated synthetic-to-instrument and instrument-to-instrument transfer, together with extension of the dynamic concentration range into the sub-ppm regime relevant to regulatory threshold monitoring, constitute the necessary next steps toward industrial deployment.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: Usage of DeepL (-Translate, -Write) and Claude (Anthropic) to improve language, clarity and fluency. After the usage, the authors reviewed and edited the content.
Conflict of interest: The authors state no conflict of interest.
Research funding: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2154 – Project ID 390874152 (POLiS Cluster of Excellence).
Data availability: The NIST reference spectra are publicly available [15]. The synthetic datasets and code are available from the corresponding author upon reasonable request.

Appendix A: Beer–Lambert superposition

Synthetic mixture spectra are assembled by additive superposition of single-compound reference spectra following Beer–Lambert law [1], [16]:

$$A(\tilde{\nu}) = l \sum_{i=1}^{K} \varepsilon_i^{\mathrm{eff}}(\tilde{\nu})\, c_i + a_{\mathrm{art}}(\tilde{\nu}), \tag{3}$$

where l = 10 cm is the optical path length, c_i is the concentration of compound i in ppm, $K \in [1, 10]$ active compounds are selected without replacement per sample and a_{art} represents additive instrumental artifacts. The effective absorption coefficients $\varepsilon_i^{\mathrm{eff}}$ are converted from the molar absorption coefficients ε_i (in L mol^{-1} cm^{-1}) via the ideal gas law at 298 K and 1 atm (0.7 % deviation from the recording temperature of 296 K, negligible in log space), giving $\varepsilon_i^{\mathrm{eff}} = \varepsilon_i / (2.445 \times 10^7)$/AU ppm^{-1} cm^{-1}.

Appendix B: Model architecture specification

Complete architecture of the chosen 1D-CNN/BNN model including SE attention, created using TensorFlow 2.20 (Table 7). Dense denotes a standard fully-connected layer.

Table 7: Full layer-by-layer specification of the 1D-CNN/BNN model. Each convolutional block follows the pattern Conv1D → BatchNorm → ReLU → MaxPool1D → SE block → AlwaysDropout. All Conv1D layers use same-padding and kernel width k = 9. SE = squeeze-and-excitation block [20] with reduction ratio r = 4. The squeeze and excitation steps use fully-connected layers with ReLU and sigmoid activations respectively. AD = AlwaysDropout with rate 0.2, active during both training and inference. Parameter counts include the non-trainable BatchNorm statistics (960 parameters). The SE model used for all reported results totals 548,720 parameters. The ablation baseline without SE blocks totals 504,600 parameters.

Component	Layer	Output shape	Param.	Notes
Input	InputLayer	(N, 3,528, 1)	0	
Block 1	Conv1D(32) + BN + ReLU	(N, 3,528, 32)	448	k = 9, same padding
	MaxPool1D	(N, 1,764, 32)	0	Stride 2
	SE block (r = 4)	(N, 1,764, 32)	552	Dense(8) → Dense(32)
	AlwaysDropout	(N, 1,764, 32)	0	Rate 0.2
Block 2	Conv1D(64) + BN + ReLU	(N, 1,764, 64)	18,752	k = 9, same padding
	MaxPool1D	(N, 882, 64)	0	Stride 2
	SE block (r = 4)	(N, 882, 64)	2,128	Dense(16) → Dense(64)
	AlwaysDropout	(N, 882, 64)	0	Rate 0.2
Block 3	Conv1D(128) + BN + ReLU	(N, 882, 128)	74,368	k = 9, same padding
	MaxPool1D	(N, 441, 128)	0	Stride 2
	SE block (r = 4)	(N, 441, 128)	8,352	Dense(32) → Dense(128)
	AlwaysDropout	(N, 441, 128)	0	Rate 0.2
Block 4	Conv1D(256) + BN + ReLU	(N, 441, 256)	296,192	k = 9, same padding
	MaxPool1D	(N, 110, 256)	0	Stride 4
	SE block (r = 4)	(N, 110, 256)	33,088	Dense(64) → Dense(256)
	AlwaysDropout	(N, 110, 256)	0	Rate 0.2
Encoder	GlobalAvgPool1D	(N, 256)	0	Spectral descriptor
BNN backbone	Dense(256) + ReLU + AD	(N, 256)	65,792	Rate 0.2
	Dense(128) + ReLU + AD	(N, 128)	32,896	Rate 0.2
Noise proxy	SpectralNoiseProxy ($\bar{\delta}$)	(N, 1)	0	From raw input
	Dense(16) + ReLU	(N, 16)	32	Noise feature
Output heads	Dense(40), linear	(N, 40)	5,160	$\hat{\mu}_i$, concentration
	Dense(40), linear	(N, 40)	5,800	$\log \hat{\sigma}^2_{\mathrm{alea},i}$; input: concat(128 ⊕ 16)
	Dense(40), sigmoid	(N, 40)	5,160	$\hat{p}_i$, presence
Total (CNN-BNN-SE, all reported results)			548,720	
Total (CNN-BNN, no SE blocks)			504,600	

Appendix C: Split conformal calibration

The normalized prediction errors $|y_{ij} - \widehat{\mu}_{ij}|/\widehat{\sigma}_{\mathrm{alea},ij}$ are computed for all n_{cal} present validation pairs, sorted in ascending order and $\widehat{q}$ is set to their empirical $(1 - \alpha)$-quantile for $\alpha = 0.05$. An appended ∞ sentinel ensures the finite-sample marginal coverage guarantee $P(y_{\mathrm{new}} \in \mathrm{PI}) \geq 1 - \alpha$ holds for exchangeable calibration pairs [25]. Present pairs within the same spectrum are correlated. Treating them as approximately exchangeable is standard practice in multi-output conformal regression. Because the bound is marginal, aggregate coverage on the test set is 0.946 while conditional coverage varies: 0.990 in the lower two thirds of each compound's concentration range, 0.859 in the upper third (0.714 in the top decile) and 0.916–0.991 across compounds. The calibration split was also used for checkpoint selection, which introduces a further deviation from exchangeability. The bound does not apply to the OOD splits or to spectra from other instruments.

Appendix D: Feature-space distance score

The feature-space distance score follows the Mahalanobis approach of Lee et al. [26]. Per-sample scores are computed as

$$D_i = \sqrt{(\mathbf{f}_i - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{f}_i - \boldsymbol{\mu})}, \tag{4}$$

where $\mathbf{f}_i$ is the 256-dimensional encoder feature vector of the CNN-BNN-SE model, averaged over 10 Monte Carlo passes. Features are first L2-normalized and projected to 64 dimensions via Principal Component Analysis (PCA) to reduce the conditioning problem before covariance estimation. Ledoit-Wolf shrinkage [27] is applied to regularize $\boldsymbol{\Sigma}$ toward a scaled identity, stabilizing the matrix inverse for high-dimensional inputs with limited calibration samples. Both the PCA projection and $\boldsymbol{\Sigma}$ are fitted on 20,000 randomly drawn in-domain test spectra and are not updated at inference time.

References

[1] J. P. Coates, "Infrared spectroscopy for process analytical applications," in *Process Analytical Technology*, Chichester, UK, John Wiley & Sons, Ltd, 2010, Chap. 6, pp. 157–194.

[2] U. Food and D. Administration, *Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, Tech. rep., U.S. Department of Health and Human Services, FDA, 2004.

[3] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Anal. Chim. Acta*, vol. 185, pp. 1–17, 1986.

[4] H. Pottel, "The use of partial least squares (PLS) in quantitative FTIR: Determination of gas concentrations in smoke gases of burning textiles," *Fire Mater.*, vol. 19, no. 5, pp. 221–231, 1995.

[5] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. Buydens, and E. Marchiori, "Convolutional neural networks for vibrational spectroscopic data analysis," *Anal. Chim. Acta*, vol. 954, pp. 22–31, 2017.

[6] D. Antonio *et al.*, *Assessing the Performance of 1D-Convolution Neural Networks to Predict Concentration of Mixture Components from Raman Spectra*, 2023, arXiv:2306.16621 [eess.SP].

[7] J. Schuetzke, N. J. Szymanski, and M. Reischl, "Validating neural networks for spectroscopic classification on a universal synthetic dataset," *npj Comput. Mater.*, vol. 9, no. 1, p. 100, 2023.

[8] H. H. Höfer, A. Orth, S. Schweidler, B. Breitung, J. Aghassi-Hagmann, and M. Reischl, "Quantitative convolutional neural network based multi-phase XRD pattern analysis," *Current Directions Biomed. Eng.*, vol. 10, no. 4, pp. 307–310, 2024.

[9] A. Orth, H. Höfer, A. Nefedov, M. Jalali, C. Wöll, and M. Reischl, "ML-based XPS quantification supported by synthetic dataset generation," *Current Directions Biomed. Eng.*, vol. 10, no. 4, pp. 482–485, 2024.

[10] R. Luo, J. Popp, and T. Bocklitz, "Deep learning for Raman spectroscopy: A review," *Analytica*, vol. 3, no. 3, pp. 287–301, 2022.

[11] P. Mishra *et al.*, "Deep learning for near-infrared spectral data modelling: Hypes and benefits," *TrAC Trends Anal. Chem.*, vol. 157, p. 116804, 2022.

[12] L. Rettenberger *et al.*, "Uncertainty-aware particle segmentation for electron microscopy at varied length scales," *npj Comput. Mater.*, vol. 10, no. 1, 2024, https://doi.org/10.1038/s41524-024-01302-w.

[13] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *Proceedings of the 33rd International Conference on Machine Learning, PMLR 48*, pp. 1050–1059, 2016.

[14] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems*, vol. 30, Red Hook, NY, USA, Curran Associates, Inc., 2017.

[15] P. M. Chu, F. R. Guenther, G. C. Rhoderick, and W. J. Lafferty, "The NIST quantitative infrared database," *J. Res. Natl. Inst. Stand. Technol.*, vol. 104, no. 1, pp. 59–81, 1999.

[16] P. R. Griffiths and J. A. De Haseth, *Fourier Transform Infrared Spectrometry*, Hoboken, NJ, USA, John Wiley & Sons, Ltd, 2007.

[17] J. Goldschmidt, L. Nitzsche, S. Wolf, A. Lambrecht, and J. Wöllenstein, "Rapid quantitative analysis of IR absorption spectra for trace gas detection by artificial neural networks trained with synthetic data," *Sensors*, vol. 22, no. 3, 2022, https://doi.org/10.3390/s22030857.

[18] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, "Enhancing deep-learning training for phase identification in powder X-ray diffractograms," *IUCrJ*, vol. 8, no. 3, pp. 408–420, 2021.

[19] X. Zhang *et al.*, "Understanding the learning mechanism of convolutional neural networks in spectral analysis," *Anal. Chim. Acta*, vol. 1119, pp. 41–51, 2020.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018. https://doi.org/10.1109/CVPR.2018.00745.

[21] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances*

in Neural Information Processing Systems, vol. 30, Red Hook, NY, USA, Curran Associates, Inc., 2017.
[22] M. Seitzer *et al.*, *On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks*, 2022, arXiv:2203.09168 [cs.LG].
[23] European Committee for Standardization (CEN), *Stationary Source Emissions – Quality Assurance of Automated Measuring Systems*, European Standard EN 14181:2014. Supersedes EN 14181:2004, European Committee for Standardization (CEN), 2014.
[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[25] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.
[26] K. Lee, *et al.*, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, vol. 31, Red Hook, NY, USA, Curran Associates, Inc., 2018.
[27] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

Bionotes



André Orth
Karlsruher Institute for Technology (KIT), Institute for Automation and Applied Informatics (IAI), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Baden-Württemberg, Germany
andre.orth@kit.edu
https://orcid.org/0009-0006-1793-6026

André Orth received the M.Sc. degree in mechanical engineering from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2021. He is a research associate with the "Automation for Laboratories" Group at the Institute for Automation and Applied Informatics (IAI), KIT. His research interests cover machine learning for the analysis of X-ray photoelectron spectroscopy (XPS) and X-ray diffraction (XRD) data, simulation-based synthetic training data generation, and uncertainty quantification.



Hawo H. Höfer
Karlsruher Institute for Technology (KIT), Institute for Automation and Applied Informatics (IAI), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Baden-Württemberg, Germany

Hawo H. Höfer received the M.Sc. degree in mechanical engineering from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2023. He is a research associate with the "Automation for Laboratories" Group at the Institute for Automation and Applied Informatics (IAI), KIT. His research interests cover machine-learning-based quantitative phase analysis of X-ray diffraction (XRD) patterns, high-throughput materials characterization, and synthetic data pipelines.



Markus Reischl
Karlsruher Institute for Technology (KIT), Institute for Automation and Applied Informatics (IAI), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Baden-Württemberg, Germany

Markus Reischl received the Dipl.-Ing. and the Ph.D. degree in mechanical engineering from the University of Karlsruhe, Germany, in 2001 and 2006 respectively. Since 2020, he has been an Adjunct Professor at the Faculty of Mechanical Engineering. He is the head of research field "Automation for Laboratories" and heading the research group "Machine Learning for High-Throughput and Mechatronics" at the Institute for Automation and Applied Informatics at the Karlsruhe Institute of Technology. Research Interests: Man-machine interfaces, image processing, machine learning, data analytics.