ARTICLE Artificial Intelligence

OPEN ACCESS

PAQC: A hybrid-intelligence framework for automated quality control of Raman spectroscopy via physics-aware and deep feature fusion

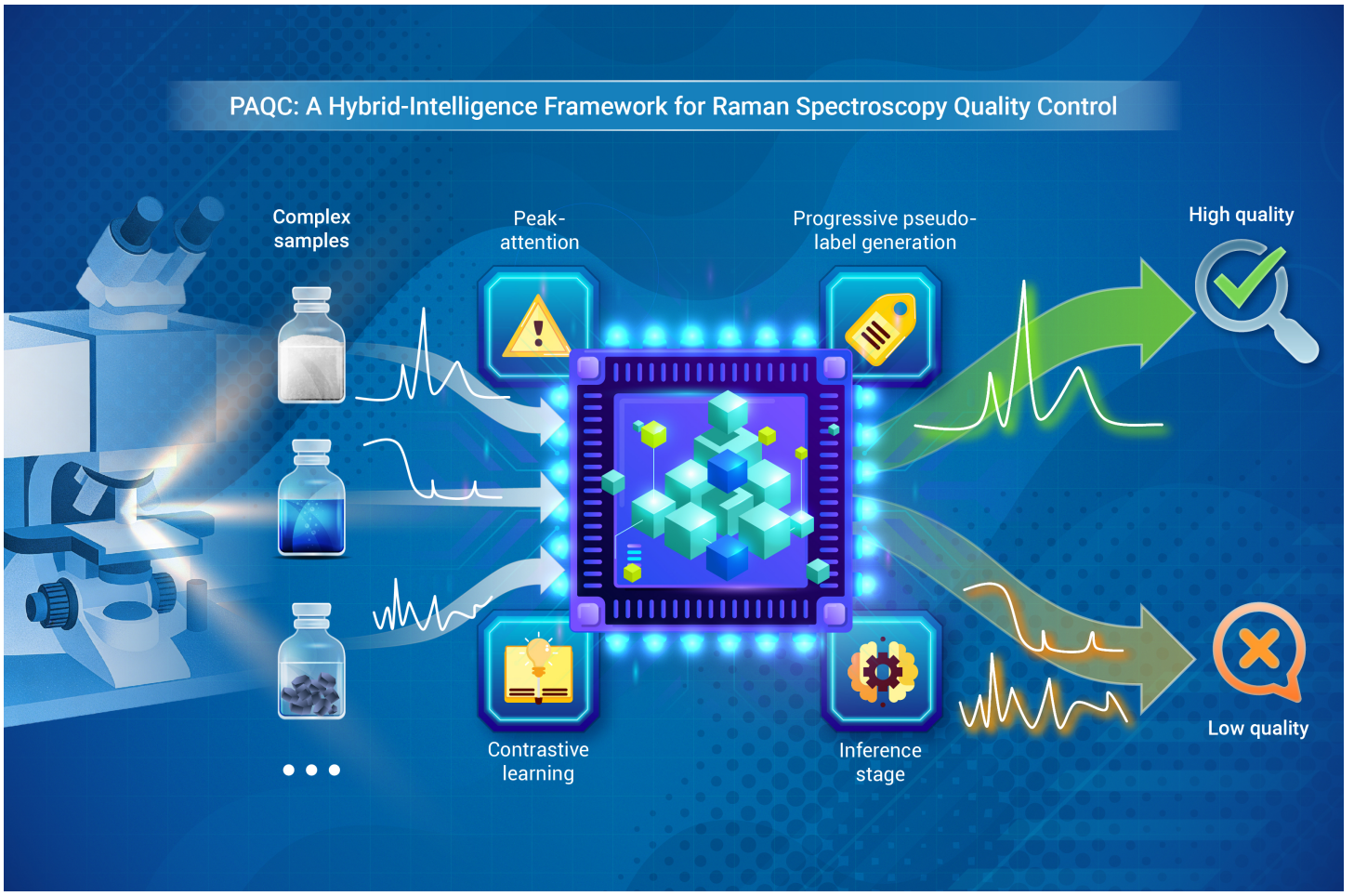
Xin Chen,[1,3] Zilong Wang,[1,3] Yuting Li,[1] Liang Li,[1] Biao Sun,[2,*] and Pei Liang[1,*]

*Correspondence: sunbiao@tju.edu.cn (B.S.); plianghust@gmail.com (P.L.)

Received: September 23, 2025; Accepted: January 5, 2026; Published Online: January 19, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100028

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Raman spectroscopy reliability is often compromised by artifacts like noise and signal saturation.
- We propose a hybrid framework synergizing explicit physical knowledge with implicit deep learning.
- This approach generates high-confidence pseudo-labels to guide model training without manual input.
- It achieves state-of-the-art accuracy and perfect recall, enabling robust automated quality control.

OPEN ACCESS

Artificial Intelligence ARTICLE

The Innovation Informatics

PAQC: A hybrid-intelligence framework for automated quality control of Raman spectroscopy via physics-aware and deep feature fusion

Xin Chen,[1,3] Zilong Wang,[1,3] Yuting Li,[1] Liang Li,[1] Biao Sun,[2,*] and Pei Liang[1,*]

[1]College of Optical and Electronic Technology, China Jiliang University, Hangzhou 310018, China

[2]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

[3]These authors contributed equally

*Correspondence: sunbiao@tju.edu.cn (B.S.); plianghust@gmail.com (P.L.)

Received: September 23, 2025; Accepted: January 5, 2026; Published Online: January 19, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100028

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Chen X., Wang Z., Li Y., et al. (2026). PAQC: A hybrid-intelligence framework for automated quality control of Raman spectroscopy via physics-aware and deep feature fusion. *The Innovation Informatics* **2**:100028.

Raman spectroscopy is a critical analytical technique across numerous engineering disciplines; however, the reliability of its measurements is often compromised by artifacts stemming from complex instrumental and sample-specific variations. Automating quality assessment poses a significant challenge: conventional physics-based thresholds, though interpretable, lack the flexibility to accommodate diverse anomalies, while data-driven deep learning approaches typically overlook valuable domain knowledge. To address this gap, we propose PAQC—a hybrid-intelligence method that synergistically integrates explicit physical priors with implicit representation learning. Our approach processes unlabeled spectral data through dual parallel streams: an explicit knowledge stream that computes physically-informed quality metrics to produce high-confidence pseudo-labels, and an implicit knowledge stream that uses a peak-attention autoencoder, guided by known peak locations, to extract discriminative deep features. These two streams are cohesively fused through a progressive contrastive learning network, yielding a highly separable feature space tailored for anomaly detection. For practical deployment, a Mahalanobis distance-based classifier enables real-time quality diagnosis of individual spectra. Evaluated on a real-world dataset using a rigorous 5-fold cross-validation protocol, PAQC achieves state-of-the-art performance with an F1-score of 98.98% ± 0.67%, while critically maintaining a perfect 100.00% recall, underscoring its effectiveness as a robust and scalable solution for automated quality control in knowledge-sensitive engineering applications.

INTRODUCTION

Raman spectroscopy, as a non-destructive molecular vibrational characterization technique, is an indispensable analytical tool in materials science, biomedicine, chemical analysis, and cultural heritage authentication since its discovery in 1928. Over the past two decades, significant advancements in laser technology, photoelectric detectors, and spectral acquisition systems have driven the development of portable Raman instrumentation, enabling a transition from benchtop equipment to handheld devices with powerful on-site detection capabilities.

During practical spectral acquisition, however, the quality and reliability of the measurement data are frequently compromised by anomalous phenomena originating from instrumental limitations and sample characteristics.[1] These anomalies, such as signal saturation, excessive noise, and strong fluorescence backgrounds, can corrupt critical spectral features, thus undermining the validity of any subsequent engineering decision-making. With the widespread implementation of chemometrics, the influence of input spectral quality on model performance has become increasingly significant.[2] As Raman technology is increasingly deployed in high-throughput and automated application scenarios, the need for a robust, objective, and scalable quality control method has become a critical challenge in engineering informatics.

Addressing this challenge automatically reveals a fundamental dilemma. Traditional physics-based models offer interpretability but often lack the flexibility to handle the complexity of real-world artifacts. In contrast, modern data-driven methods provide powerful feature learning but typically operate as "black boxes" that overlook valuable domain knowledge and, more critically, require large-scale curated data—creating a practical paradox.

The first paradigm, rooted in physics, addresses quality control by quantifying explicit, known artifacts through interpretable metrics. These metrics target common anomalies such as spectral saturation,[3] where intense signals exceed the detector's dynamic range, and baseline distortion, often caused by sample fluorescence or instrumental effects.[4–8] They also assess spectral noise arising from factors like photon shot noise and detector thermal noise, typically using variations of the Signal-to-Noise Ratio (SNR).[9–16] While these physics-based indicators provide an explicit representation of quality, they often struggle to capture the complex interplay between different artifacts and may require task-specific tuning, limiting their generalizability when used in isolation as a complete quality control solution.

The second paradigm leverages purely data-driven algorithms to learn patterns of "normal" and "anomalous" behavior directly from the data, without explicit knowledge of the artifact types. The evolution of these methods began with classical statistical approaches like Principal Component Analysis (PCA)[17] and Isolation Forest,[18] which often struggle with the high dimensionality and complex non-linear structures inherent in spectral data. While shallow models such as the One-Class Support Vector Machine (OC-SVM) offered theoretical improvements by modeling only the normal data distribution,[19–22] they still face significant challenges in constructing a robust feature space for diverse spectra. Consequently, deep learning has emerged as a powerful alternative, with architectures based on autoencoders[23] and Deep Support Vector Data Description (Deep SVDD)[24–27] demonstrating far superior feature learning capabilities. Despite their architectural diversity, however, these advanced approaches universally converge on a critical bottleneck: their performance relies heavily on large-scale, pre-screened, high-quality normal training data. This creates a practical paradox: to build an effective anomaly detector, one must first manually perform the very task of anomaly detection to curate a clean training set. This expensive and often subjective pre-processing step hinders scalability and creates a fundamental conflict between model robustness and data acquisition demands, highlighting the urgent need for a new paradigm capable of learning from un-curated, real-world engineering data.

To overcome the limitations of purely physics-based and purely data-driven models, as discussed in the Introduction, the development of hybrid-intelligence methods has become a significant trend in the field of engineering informatics. The core idea lies in combining the interpretability afforded by domain knowledge with the powerful non-linear fitting capabilities of deep learning to build more robust and trustworthy engineering decision-support systems. This paradigm has seen wide application in the engineering domain. For instance, Wei et al.[28] and Sun et al.[29] in their work utilized prior physical knowledge to generate pseudo-labels, successfully applying them to tasks such as fault detection. A more advanced direction involves directly integrating physics information (such as partial differential equations) as a constraint into the neural network's training process, known as Physics-Informed Neural Networks (PINNs). This technique has also shown great potential in fields like fault diagnosis.[30,31]

The PAQC framework proposed in this work adheres to this hybrid-intelligence paradigm but is distinguished by the depth and strategy of its knowledge fusion compared to existing research. Rather than simply generating static pseudo-labels, PAQC introduces a novel progressive curriculum learning strategy that dynamically manages uncertainty. This study focuses on three global anomaly types that are paramount for automated quality control: signal saturation, elevated baseline and strong noise. While localized anomalies

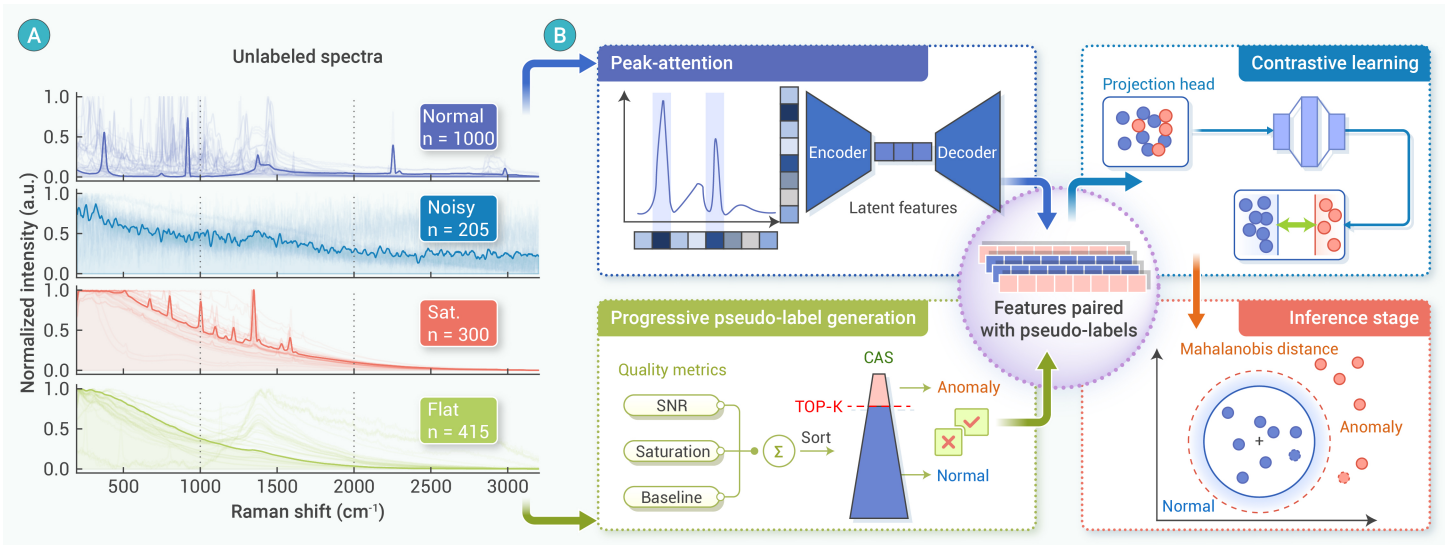


Figure 1. Data Challenges and System Architecture of PAQC framework (A) Representative raw Raman spectra illustrating the four primary categories in the dataset: Normal (high-quality), Noisy (low SNR), Saturated (signal clipping), and Flat (fluorescence interference). (B) The overall workflow of the proposed PAQC hybrid-intelligence framework. It processes unlabeled spectra through two parallel streams: an explicit knowledge stream calculating physics-informed metrics (SI, SBI, NI) to generate high-confidence pseudo-labels, and an implicit knowledge stream utilizing a Peak-Attention Autoencoder to extract deep features. These are fused via a progressive contrastive learning.

like cosmic ray spikes are important, our primary goal is to first assess the foundational quality of a measurement, a critical step for any robust instrumentation system. The PAQC method processes unlabeled data through two parallel streams: (1) an explicit knowledge stream employs a physics-informed evaluation module to generate high-confidence pseudo-labels; (2) concurrently, an implicit knowledge stream utilizes a peak-attention autoencoder, which leverages explicit peak-location priors, to extract robust deep features. The outputs are then fused in a progressive contrastive learning network to build a highly accurate diagnostic model, which is deployable for real-time, single-spectrum classification.

The main contributions of this work are summarized as follows:

(1) We propose a novel hybrid-intelligence paradigm for engineering signal quality assessment that effectively resolves the conflict between traditional physics-based models and data-driven black-box models.

(2) We develop a new technique for knowledge fusion where explicit, physics-informed metrics are used to generate reliable supervisory signals that guide a self-supervised contrastive learning network.

(3) We design and validate an end-to-end automated method on a real-world dataset, demonstrating its state-of-the-art performance and its potential as a robust quality control solution for knowledge-intensive engineering applications.

MATERIALS AND METHODS

Dataset and preprocessing

To rigorously evaluate the PAQC framework under realistic conditions, we compiled a comprehensive dataset derived from authentic experimental measurements. This dataset was not artificially generated but was accumulated over the course of numerous routine experiments, ensuring it encapsulates the diversity and unpredictability of real-world data acquisition. It comprises measurements from a variety of different samples under a range of testing conditions, such as varied laser power, integration times, and sample positioning. This process naturally produced spectra with a wide spectrum of quality levels. All 1920 spectra were collected using a portable Raman spectrometer, covering a wavenumber range of 202-3200 cm^{-1} with 1500 data points per spectrum. Each spectrum was then manually inspected and labeled by experts, yielding four distinct categories that represent authentic measurement outcomes:

Normal (1000 samples) with clear peaks and stable baselines; and three types of anomalous spectra corresponding to common experimental artifacts—Noisy (205 samples) with significant random fluctuations, Saturated (300 samples) with peak clipping, and Flat (415 samples) with overwhelming fluorescence background. Representative examples of each category are shown in Figure 1A. To preserve the original artifact characteristics for our unsupervised task, the only preprocessing applied was min-max normalization.

For a rigorous evaluation of the model's performance and to ensure the generalizability of our findings, the complete dataset of 1920 spectra was partitioned using a 5-fold cross-validation (CV) strategy. This protocol was adopted to provide a robust estimation of performance and directly address the limitations of relying on a single data partition. The dataset was first shuffled and divided into 5 non-overlapping folds of equal size. In each of the 5 iterations, 4 folds were used as the training set (for both autoencoder pre-training and contrastive learning) and the remaining fold was held out as the test set. The offline calibration step was performed using the 'Normal' samples identified within the 4-fold training set. This process was repeated 5 times, ensuring that every spectrum was used for testing exactly once. All performance metrics reported in this manuscript (e.g., Table 1) are the mean ± standard deviation ($mean \pm std$) across these 5 independent runs. This protocol provides a robust and unbiased estimate of the model's true generalization capability.

System overview

The overall architecture of the PAQC framework, illustrated in Figure 1B,is designed around two parallel processes. In the first, a series of physics-informed metrics are calculated for each spectrum to produce a Comprehensive Anomaly Score (CAS). These scores are then converted into high-confidence pseudo-labels. Concurrently, a Peak-Attention Autoencoder learns a robust, low-dimensional feature representation from the raw spectral data. Finally, the outputs from these two processes—the deep features and the physics-guided pseudo-labels—are integrated within a progressive contrastive learning network to optimize the feature space for precise anomaly detection.

Physics-informed quality metrics

The cornerstone of the PAQC framework is the use of physics-informed quality metrics. This module's task is to formalize established domain knowledge from spectroscopy into a series of computable, interpretable physical indicators, enabling a rapid, coarse-grained quality assessment. A key design choice is for these indicators to be substance-agnostic, ensuring broad applicability. We therefore focus on universal artifacts like saturation, noise, and baseline distortion.

Saturation Index (SI). The direct morphological characteristic of saturated spectra is the presence of plateau structures at peak tops, with intensities approaching maximum values. To quantitatively characterize this, a Spectral Saturation Index (SI) was developed based on a multi-threshold sliding window analysis. The calculation is based on the following framework: First, three characteristic thresholds (T_1=0.95,T_2=0.97, T_3=0.99) are estab-

lished for the normalized spectrum, representing warning, moderate, and severe saturation states, respectively. Considering the decisive role of high-threshold regions in saturation determination, different importance weights are assigned (ω_1 = 0.15, ω_2 = 0.25, ω_3 = 0.60). Second, a sliding window (size = 10, overlap = 50%) is employed for local spectral analysis. For each window w, the proportion of points $C(\omega, T_i)$ with signal intensities exceeding threshold T_i is calculated, and high-coverage windows where $C(\omega, T_i) > 0.5$ are identified. For each threshold T_i, the proportion of these high-coverage windows among all windows, $R(T_i)$, is calculated as the basic saturation characteristic. Finally, the SI is computed via a weighted summation:

$$SI = \sum_i \omega_i \cdot R(T_i) \tag{1}$$

where $R(T_i)$ is the proportion of high-coverage windows exceeding a high-intensity threshold T_i (e.g., 0.95, 0.99), and $_i$ is the corresponding weight. This formulation is designed to be particularly sensitive to "flat-top" clipping, a critical artifact. Higher weights ($_i$) are assigned to higher thresholds (T), reflecting the physical understanding that a spectrum approaching 100% intensity is a more severe and unambiguous indicator of saturation than one just approaching 95%. This index effectively quantifies the degree of signal saturation.

Spectral Baseline Index (SBI). To address the requirements for Raman spectral baseline quality evaluation, a multi-region normalized baseline area evaluation method is proposed. This method employs a region segmentation strategy, dividing the Raman spectrum into three critical regions (200-1000 cm^{-1}, 1002-2000 cm^{-1}, and 2002-3200 cm^{-1} in our dataset), corresponding to the fingerprint region, mid-frequency region, and high-frequency region, respectively, enabling fine-grained assessment of baseline characteristics across different regions. The baseline in each region is estimated using the adaptive iteratively reweighted penalized least squares (airPLS) algorithm, which is robust to the presence of Raman peaks. The Spectral Baseline Index (SBI) for each region is then calculated as the normalized area under the estimated baseline curve, making the indices comparable across regions of different widths.

$$SBI_i = \frac{1}{\Delta v_i} \int_{v_{i,min}}^{v_{i,max}} B_i(v)\, dv \tag{2}$$

where SBI_i represents the normalized baseline area index within the i-th wavenumber interval, $B_i(v)$ represents the baseline function extracted via airPLS method; v represents Raman shift in cm^{-1}; $v_{i,max}$ and $v_{i,min}$ represent the maximum and minimum wavenumbers of the i-th interval, respectively; and v_i represents the width of the wavenumber interval, i.e., $v_i = v_{i,max} - v_{i,min}$.

Noise Index (NI). This research employs a noise calculation approach based on baseline removal. The method first applies the aforementioned airPLS algorithm for baseline correction. Subsequently, an adaptive threshold peak detection strategy is adopted. The whole-spectrum standard deviation serves as an initial (typically overestimated) noise level estimate, and the detection threshold is set at twice this level to ensure that low-intensity noise is filtered while retaining genuine characteristic peaks. For the SNR calculation, peak attributes are utilized to construct masks, dividing peak and non-peak regions. For background noise estimation, the Median Absolute Deviation (MAD) is employed instead of the traditional standard deviation to enhance robustness against outliers. The SNR is calculated as the ratio of the average peak intensity to the background noise level, and converted to a decibel representation:

$$\mathrm{SNR_{dB}} = 10 \cdot \log_{10}\left(\frac{I_{signal}}{I_{noise}}\right) \tag{3}$$

$$I_{signal} = \frac{1}{N_p}\sum_{i=1}^{N_p} I_{peak,i}; I_{noise} = \mathrm{median}\left(|I_{bg} - \mathrm{median}(I_{bg})|\right)$$

where I_{signal} represents the average value of all N_p detected Raman peak intensities, and I_{noise} represents the Median Absolute Deviation (MAD) of the background region signal separated by masking. To facilitate integration with other indicators, the SNR (in dB) is converted into a Noise Index (NI) within the [0,1] range via a piecewise linear mapping function. This mapping is based on empirical thresholds for Raman spectral quality: an NI of 1.0 (extremely high noise) is assigned to SNR ≤ 5 dB, and 0.0 (negligible noise) to SNR ≥ 30 dB, with linear variation in between. This normalization strategy enables the noise indicator to be weighted and combined with other indicators, establishing a foundation for subsequent comprehensive anomaly scoring.

$$NI = \max\left(0, \min\left(1, 1 - \frac{SNR_{dB} - 5}{25}\right)\right) \tag{4}$$

Comprehensive anomaly score calculation. The individual indices are integrated into a single Comprehensive Anomaly Score (CAS) via a weighted linear combination, providing a holistic measure of spectral quality.

$$CAS = \sum_{i=1}^{N_m} w_i \cdot I_i \tag{5}$$

where I_i denotes the i-th evaluation index (SI, SBI, NI) and w_i is its corresponding weight, and N_m is the total number of metrics (in this study, $N_m = 3$). A key advantage of this model is its flexibility: the weights w_i can be tuned to adapt to specific application scenarios. For instance, in analytical tasks highly sensitive to peak shape, the weight for the Saturation Index (SI) could be increased, whereas for samples with strong fluorescence backgrounds, the weight for the Spectral Baseline Index (SBI) could be prioritized.

For reproducibility in this study, the weights were set based on empirical analysis of the artifacts' impact (SI = 0.3, NI = 0.5, SBI = 0.2), prioritizing the noise component. A detailed justification for these physics-informed parameters is provided in the Supplemental Information (Section S1.3). This computationally efficient scoring provides a robust foundation for the subsequent pseudo-label generation. Detailed validation and visual analysis for each of these three physical indices are presented in the RESULTS section, demonstrating their effectiveness in distinguishing different spectral categories.

The specific values for these physics-informed parameters (e.g., SI thresholds, peak detection prominence, G_{peak} smoothing sigma) are not treated as data-driven hyperparameters but as codified domain knowledge based on instrument characteristics and the physical definition of spectral features. A detailed justification for these parameter choices in the context of their physical meaning is provided in the Supplemental Information (Section S1.3).

Physics-guided feature extraction

The first part of our implicit knowledge engine is a deep autoencoder designed not as a black box, but as a feature extractor that is actively guided by physical priors. While a standard autoencoder treats all parts of the spectrum equally, our proposed Peak-Attention Enhanced Autoencoder focuses its representational power on the physically meaningful characteristic peaks—the primary carriers of chemical information (Figure 2).

The core of this guidance is a physics-informed guidance matrix, G_{peak}, which incorporates prior knowledge about the nature of Raman spectra. The generation of G_{peak} follows a multi-step process: First, a high prominence threshold (prominence > 0.3) is established for peak detection, effectively filtering weak peaks and noise to retain only principal peaks with significant physical meaning. Then, a binary guidance matrix is constructed at the identified peak positions. Subsequently, a Gaussian filter is applied to smooth these regions, creating a gradient attention weight distribution that simulates the half-width characteristic of Raman peaks. Finally, the matrix undergoes normalization to ensure its weight distribution is within the [0,1] range.

To efficiently integrate this physical prior, the model adopts a dual-channel attention architecture that processes both local peak features and global spectral information. The peak detection channel comprises three convolutional neural network layers to capture peak morphological features, while the feature attention channel adopts a multilayer perceptron for global context. The outputs from both channels, A_{peak} and $A_{feature}$, are dynamically fused through learnable parameters α and β. These parameters are processed through a sigmoid function and then normalized to ensure their sum equals 1, realizing an adaptive balance between the two attention components:

$$A_{combined} = \alpha \cdot A_{peak} + \beta \cdot A_{feature} \tag{6}$$

The final attention weights are updated according to a fixed proportion:

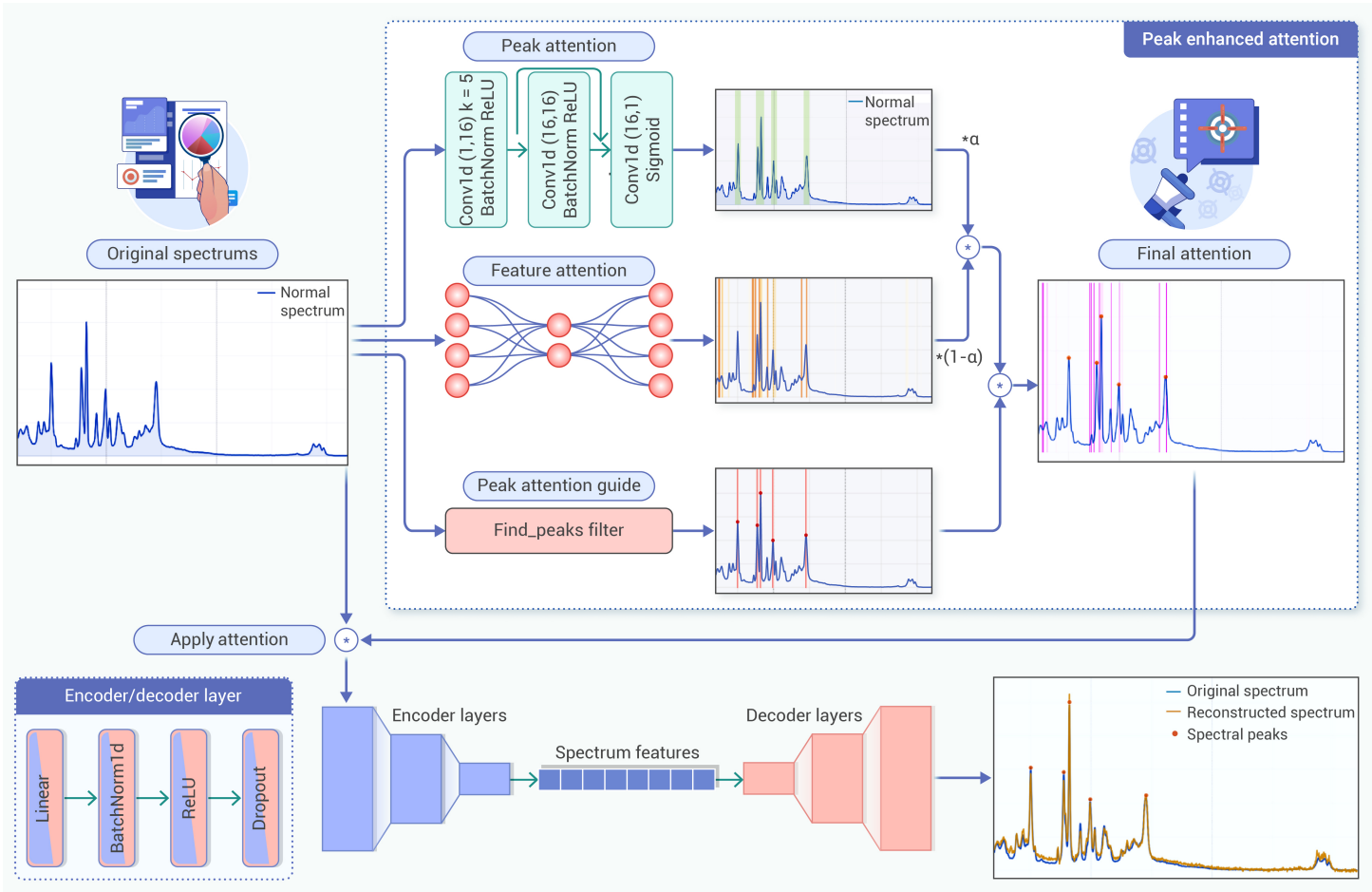


Figure 2. Architecture of the Peak-attention enhanced Autoencoder.

$$A_{final} = 0.7 \cdot (A_{combined} \cdot G_{peak}) + 0.3 \cdot A_{combined} \quad (7)$$

where G_{peak} represents the peak guidance matrix. This fusion strategy strongly biases the network's focus towards the physically significant regions identified by G_{peak}, while the $0.3 * A_{combined}$ term preserves a degree of data-driven exploration, preventing over-reliance on the physical prior. This ensures that the feature extraction process is both anchored by established knowledge and adaptive to new patterns.

A complete, detailed description of this dual-channel architecture, including the specific layer dimensions, residual connections, and activation functions, is provided in the Supplemental Information (Section S1.1.1)

Progressive pseudo-label generation

This stage serves as the critical bridge between the physics-informed evaluation and the deep representation learning. The continuous Comprehensive Anomaly Scores (CAS), derived from the physics-informed metrics (Eq. 5), are converted into discrete pseudo-labels. These labels provide the crucial supervisory signal for the subsequent contrastive learning stage. This conversion process is designed to be both robust and adaptive.

Instead of using a fixed threshold, which is brittle across different data batches, we employ a dynamic, ranking-based strategy. For each batch, spectra are ranked by their CAS scores. The top k% of samples with the highest scores are assigned the 'Anomaly' pseudo-label, while the remaining (100-k)% are designated as 'Normal'. This relative, top-k approach ensures robustness against distributional shifts between batches.

To further enhance learning stability, we introduce a progressive, curriculum-learning-inspired strategy. The anomaly ratio k is dynamically adjusted across training epochs. Based on a prior estimation that the true anomaly ratio in our dataset is approximately 15%, we designed a three-stage curriculum:

Initial Stage (k = 10%): The model is trained on only the most confident, unambiguous anomalies, establishing a stable initial feature space.

Middle Stage (k = 15%): The anomaly ratio is increased to match the estimated true proportion, training the model on a well-balanced set of samples to learn finer-grained distinctions.

Final Stage (k = 20%): The model is challenged with the most difficult, ambiguous boundary cases by extending the ratio beyond the estimated true proportion, refining its final discriminative power.

This progressive strategy effectively balances the trade-off between pseudo-label quality and quantity, mitigating the impact of potential label noise in early training and ensuring a smooth, robust optimization process.

Guidelines for Unknown Anomaly Rates: For real-world scenarios where the ground-truth anomaly rate is unknown, we propose a "Distribution-Driven Heuristic" to select *k*, underpinned by the framework's inherent robustness via decoupled design.

1. Distribution-driven heuristic: Operators should first analyze the histogram of CAS scores for a calibration batch (as shown in Figure 6C). We recommend setting the target *k* at the "elbow point" of the distribution or the valley trailing the dominant normal peak. This statistical natural boundary provides a data-driven, logical estimation of the anomaly proportion.

2. Inherent robustness via decoupled design: Crucially, the framework is designed to be insensitive to the precision of this selection. This robustness stems from the decoupling of representation learning and inference: the parameter *k* serves only to structure the feature space, while the final decision rigor is independently governed by the inference threshold (e.g., 99.0th percentile). Furthermore, the progressive curriculum prevents label noise from corrupting the feature space even if *k* is overestimated. This dual mechanism ensures that a simple, approximate choice of *k* does not compromise the final diagnostic performance.

Pseudo-label guided representation learning

The final stage of the PAQC framework uses the supervisory signal generated from the physics-informed metrics to guide the final optimization of the

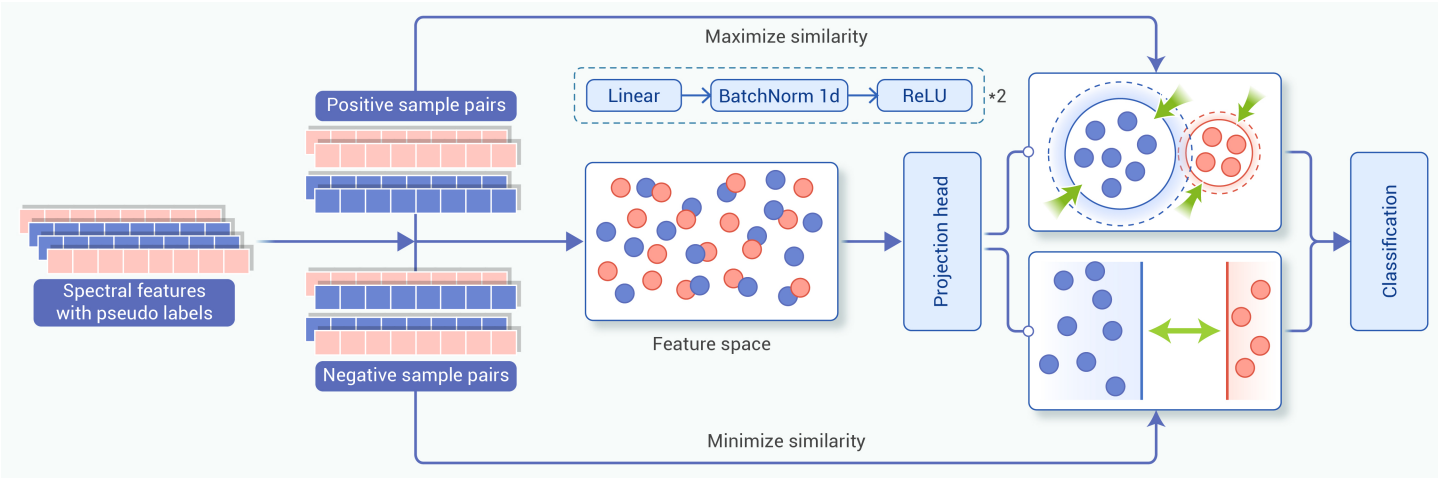


Figure 3. Schematic diagram of the contrastive learning strategy.

deep feature representation. This section details the contrastive learning strategy used to learn this discriminative representation and the efficient inference mechanism for real-time diagnosis.

The contrastive learning network utilizes the features from the peak-guided autoencoder as input, projecting them into a new space via a non-linear projection head. The projection head is a three-layer MLP with a "narrow-wide-narrow" architecture (input dimension → 256 → 128 → 64), incorporating batch normalization and ReLU activation. This structure enhances the model's expressive capability while preserving discriminative information. To generate positive pairs for contrastive learning, two different augmented views are created for each sample by applying random noise (at a 0.05 level) and slight frequency shifts.

Our core innovation lies in how the supervisory signal derived from the explicit knowledge component is used to structure the learned feature space. Unlike traditional contrastive learning that only treats augmented views of the same sample as positive pairs, our approach introduces class-level consistency defined by the pseudo-labels. As illustrated in Figure 3, positive pairs are formed from two mechanisms: (1) self-augmentation consistency, where different views of the same sample are paired; and (2) class consistency, where different samples sharing the same pseudo-label are paired. Conversely, samples with different pseudo-labels form negative pairs. This dual-constraint mechanism, optimized via a contrastive loss function (Eq. 8), which enables the model to learn a highly discriminative feature representation. For any given anchor sample z_i, the loss is designed to maximize the similarity with its positive pairs$P(i)$ (samples with the same pseudo-label) relative to all other samples in the batch:

$$L_i = -\log \frac{\sum_{z_p \in P(i)} \exp(\mathrm{sim}(z_i, z_p)/)}{\sum_{z_a \in A(i)} \exp(\mathrm{sim}(z_i, z_a)/)} \quad (8)$$

where: $P(i)$ is the set of all positive pairs for z_i in the batch (including z_i itself or its augmentation). $A(i)$ is the set of all samples in the batch (positives and negatives). is the temperature parameter, set to 0.3 as specified in Section S1.2.4.

For deployment and inference, we adopt an efficient real-time diagnostic strategy. This process is divided into two phases:

1. Offline Calibration: After the contrastive learning network is fully trained, all high-confidence 'Normal' samples from the training set are passed through the frozen feature extractor. The distribution of these normal features is then modeled by calculating its mean vector ($_{\text{normal}}$) and its regularized inverse covariance matrix ($\sum^{-1}_{\text{norma}}$) in the feature space.

2. Online Diagnosis: For any new, incoming single spectrum, the system performs a rapid diagnosis. The spectrum is first passed through the fixed feature extractor to obtain its feature vector. Then, its Mahalanobis distance to the pre-calibrated 'Normal' centroid is calculated. By comparing this distance to a pre-defined threshold, an instantaneous and robust classification of the spectrum's quality is achieved, making the framework suitable for real-time engineering applications.

Critically, the selection of this threshold is not based on a simple default percentile (e.g., 95%). Instead, our final decision threshold was determined through a comprehensive sensitivity analysis (presented in Supplemental Information, Section S4.2). This analysis identified the 99.0th percentile of the normal calibration distances as the optimal balance for a quality control application. This choice was made because it represents the optimal trade-off that achieves a state-of-the-art F1-score while simultaneously ensuring that all true anomalies are identified (i.e., maintaining perfect recall), a critical requirement for quality control systems.

RESULTS

Validation of physics-informed quality metrics

This section validates the effectiveness of the physics-informed quality metrics, which form the foundation of the PAQC framework's guidance system. Each of the three individual indices—Saturation (SI), Baseline (SBI), and Noise (NI)—is analyzed separately to confirm its ability to distinguish different spectral characteristics based on physical principles.

Figure 4A demonstrates the detection effectiveness of the saturation index, revealing significant differences in saturation characteristics among different types of spectra. Normal spectra (blue) exhibit clear characteristic peaks without any saturated regions, resulting in an SI value of 0.000. In contrast, saturated and flat spectra show obvious saturation features in the low wavenumber regions. The violin plot of the SI distribution further quantifies these characteristics across the entire dataset, showing that while Normal and Noisy samples have SI values concentrated near 0, Saturated samples have a significantly higher and wider SI distribution (median of 0.26). The adopted sliding window analysis effectively distinguishes between natural Raman peaks and saturated regions by recognizing fundamental differences in their morphology; Raman peaks exhibit natural shapes (Lorentzian or Gaussian), whereas saturated regions appear as unnatural flat-top truncations.

Figure 4B shows the multi-region baseline analysis, revealing characteristic differences that are valuable for anomaly detection. The violin plot quantifies the distribution of baseline areas, demonstrating significant differences across the three wavenumber intervals, which reflects the wavelength-dependent nature of fluorescence interference. A key distinction between Saturated and Flat spectra lies in the signal characteristics after baseline removal: Saturated spectra, despite strong fluorescence, still retain distinct Raman peaks, whereas Flat spectra show almost no identifiable peaks. This indicates different underlying anomaly mechanisms—signal truncation for saturated spectra versus a complete lack of Raman activity or overwhelming fluorescence for flat spectra.

Figure 4C shows the Noise Index (NI) analysis results. The SNR distribution plot on the left reveals significant differences among the four spectral classes. Normal spectra exhibit the highest SNR (mean μ = 22.3 dB), while Noisy spectra show an extremely low SNR (mean μ = 8.2 dB), allowing this metric to almost completely separate these two classes. However, Saturated and Flat spectra partially overlap with Normal spectra in the SNR dimension.

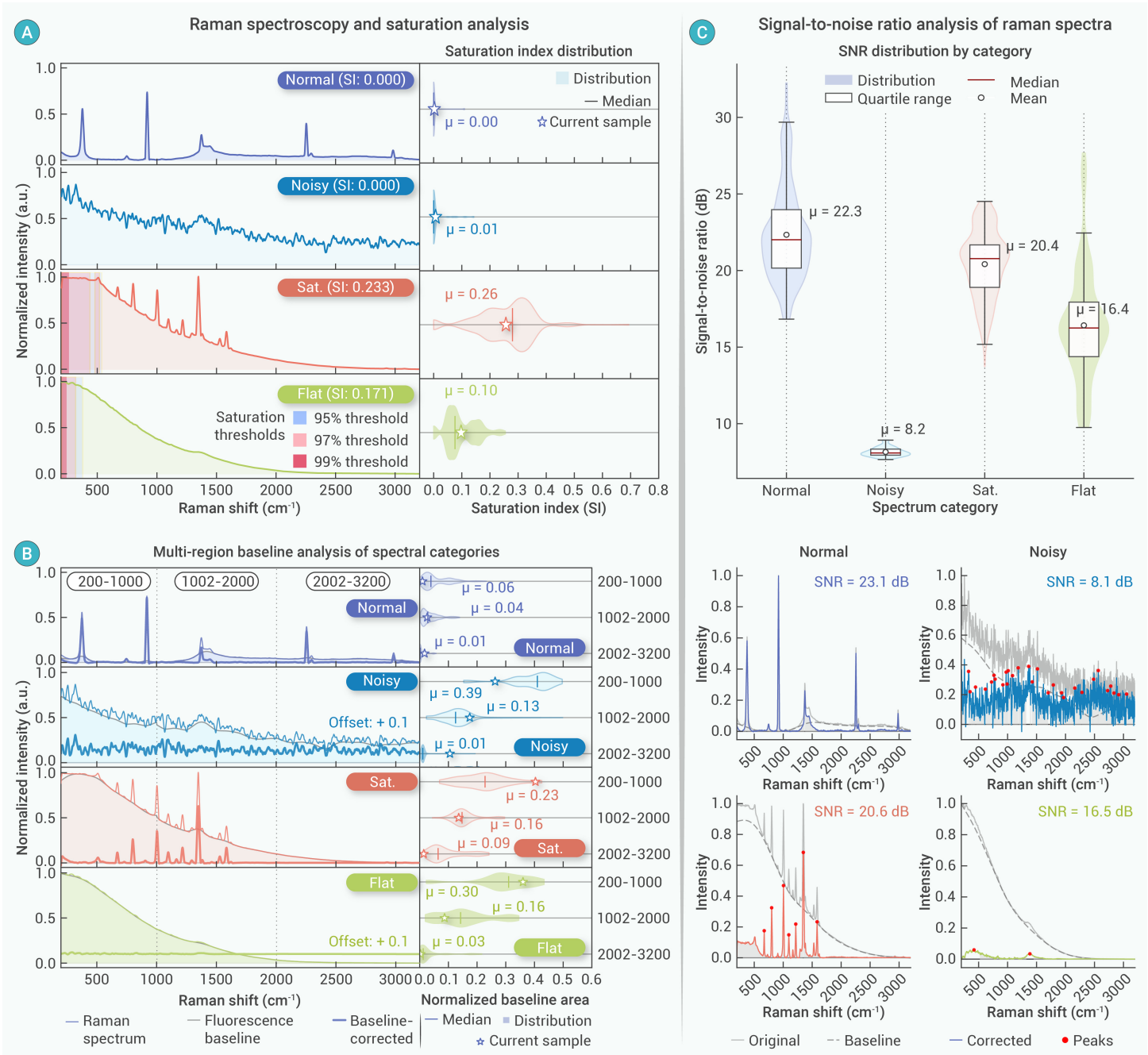


Figure 4. Visualization and statistical analysis of physics-informed quality metrics (A) Saturation Analysis (SI): Comparison of spectral morphologies (Left) shows distinct "flat-top" plateaus in saturated spectra compared to normal peaks. The SI distribution (Right) confirms significantly higher values for the Saturated class. (B) Baseline Analysis (SBI): Multi-region segmentation (200-1000, 1002-2000, 2002-3200 cm^{-1}) reveals that Flat spectra possess significantly larger baseline areas due to fluorescence. (C) Noise Analysis (NI): Signal-to-Noise Ratio (SNR) distributions demonstrate that the Noise Index derived from SNR effectively separates Noisy spectra (mean SNR ≈ 8.2 dB) from other categories.

This distribution reaffirms the difficulty of distinguishing all anomaly types with a single indicator and justifies the necessity of our multi-indicator fusion approach to capture the characteristics of different anomaly types comprehensively.

Collectively, the analyses of the Saturation, Baseline, and Noise Indices demonstrate that our physics-informed module forms a robust and well-differentiated quality scoring system. While each index effectively targets a specific type of spectral artifact, their real power lies in their combined application. Different anomaly types exhibit distinct signatures across these indices, providing a comprehensive and interpretable view of spectral quality. Crucially, this multi-indicator evaluation is computationally efficient, with an average processing time of approximately 0.03s per spectrum. This efficiency ensures that the explicit knowledge component can operate effectively in large-scale datasets, providing a rapid and robust quality assessment that lays a solid foundation for the subsequent deep learning stages of our framework.

Validation of physics-guided feature extraction

This section validates the effectiveness of the physics-guided feature extraction module. We provide a comprehensive evaluation of how using an explicit physical prior—the locations of known spectral peaks—to guide the autoencoder's attention mechanism impacts reconstruction quality. The analysis covers both spectral reconstructions and attention weight distributions, as shown in Figure 5.

Figure 5A illustrates the reconstruction comparison for normal and saturated spectra. For normal spectra, the introduction of peak guidance reduced the MSE by 19.24%. The improvement was more significant for artifact-heavy spectra like the saturated example, which saw a 97.95% reduction in MSE.

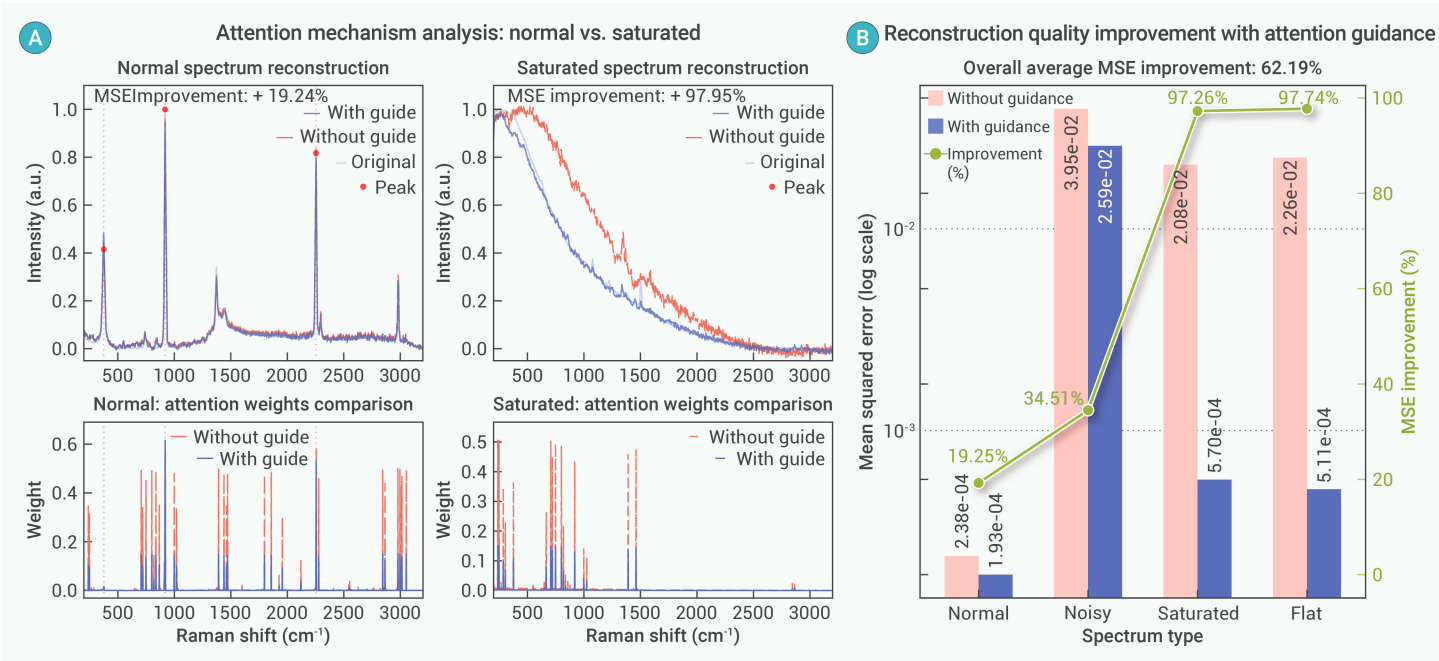


Figure 5. Analysis of the impact of Peak-Guided attention mechanism on Autoencoder reconstruction quality (A) Qualitative comparison of spectral reconstruction and attention weight distributions for Normal and Saturated spectra. The "With guide" attention weights (red lines) sharply focus on characteristic peaks, leading to precise peak reconstruction, whereas "Without guide" weights (blue lines) are dispersed. (B) Quantitative evaluation of reconstruction quality improvement. The bar chart shows the significant reduction in Mean Squared Error (MSE) across all spectral categories after introducing the peak-attention mechanism, with an overall average improvement of 62.19%.

The attention weight distribution plot below it clearly demonstrates the core function of this guidance: the attention weights are transformed from a dispersed state to one sharply concentrated on the physically significant Raman peaks. This "focused" attention enables the model to prioritize preserving the position, intensity, and morphology of Raman peaks, which are key characteristics for sample identification.

Figure 5B further quantifies the reconstruction improvements across all four spectral types. The results indicate that the mechanism brings significant quality improvements to all categories, with the most notable reductions in MSE for noisy (97.26%) and flat (97.54%) spectra. The overall average MSE improvement across all classes reached 62.19%, thoroughly validating the effectiveness of the proposed peak-guided attention mechanism.

From a mechanistic perspective, peak-guided attention enhances weights near peak points while reducing them in non-peak regions. For normal spectra, this strengthens the precise reconstruction of characteristic peaks; for anomalous spectra, it helps reduce interference from non-informative regions (e.g., noise, steep baselines). It was also observed that the model's reconstruction capability for normal spectra was significantly better than for anomalous spectra, which aligns with the basic concept of using autoencoders for anomaly detection—highlighting anomalies by optimizing reconstruction for the normal class.

In summary, this comprehensive analysis indicates that the peak-guided attention mechanism not only improves overall reconstruction quality but, more importantly, enhances the model's sensitivity to the physical characteristics of Raman spectra. It enables the implicit knowledge engine to perform representation learning based on the essential features of spectra rather than superficial noise. This establishes a solid foundation for the subsequent stages, ensuring that the model captures features that truly reflect molecular vibration information.

Analysis of the progressive pseudo-labeling strategy

This section analyzes the critical bridge between the two parts of our framework: the progressive pseudo-labeling strategy. We first examine the distribution of the anomaly scores derived from the physics-informed metrics and then visualize how the progressive, top-k% strategy translates these scores into an evolving set of supervisory labels for the subsequent learning stage, as shown in Figure 6.

Figure 6C shows the anomaly score distributions for each true class. A clear separation is visible: Normal spectra (blue) consistently receive low scores, while Saturated spectra (red) receive the highest. Noisy (yellow) and Flat (green) spectra fall in between, with some overlap. This distribution confirms that our physics-informed scores effectively map spectral artifacts to a quantitative scale, but also highlights the challenge of using a single, fixed threshold, as it would struggle with the overlapping boundary regions.

To address this, our progressive, top-k% strategy generates pseudo-labels in an "easy-to-difficult" curriculum. The t-SNE visualizations in Figure 6A (true labels) and Figure 6B (pseudo-labels) demonstrate this process. At the initial 10% threshold (dark red), the model correctly identifies the core of the most obvious anomaly cluster (Saturated spectra). As the threshold increases to 15% and 20% (lighter shades of red), the pseudo-labeling expands outwards, gradually incorporating the more ambiguous Noisy and Flat spectra located at the cluster boundaries.

It is also instructive to analyze the discrepancies between the pseudo-labels and the true labels. Notably, some samples with "Noisy" or "Flat" true labels are assigned "Normal" pseudo-labels (e.g., in the circled region of Figure 6). This is not a failure, but rather an intentional feature of our conservative, curriculum-based approach. The framework deliberately avoids assigning "Anomaly" labels to these highly ambiguous, low-score boundary cases, especially in the early stages, to maintain the purity of the supervisory signal. It prioritizes building a robust decision boundary based on high-confidence examples first.

This inherent caution is a key reason for the framework's stability. By starting with high-purity pseudo-labels and progressively introducing more complex examples, our strategy effectively mitigates the risk of the model being corrupted by label noise during early training—a common failure mode in pseudo-labeling systems. This provides a high-quality, evolving supervisory signal that is crucial for the success of the final contrastive learning stage.

Overall performance of the PAQC framework

The final performance of the integrated PAQC framework is comprehensively presented in Figure 7, which validates both the quality of the learned feature space and the effectiveness of the rigorously-defined real-time diagnostic classifier.

Overall anomaly detection performance. The analysis begins with the general structure of the feature space, as intuitively visualized in the 3D Principal Component Analysis (PCA) view (Figure 7B). This plot demonstrates that the normal (blue) and anomalous (red) samples form two highly distinct and compact clusters that are linearly separable, with the few misclassified

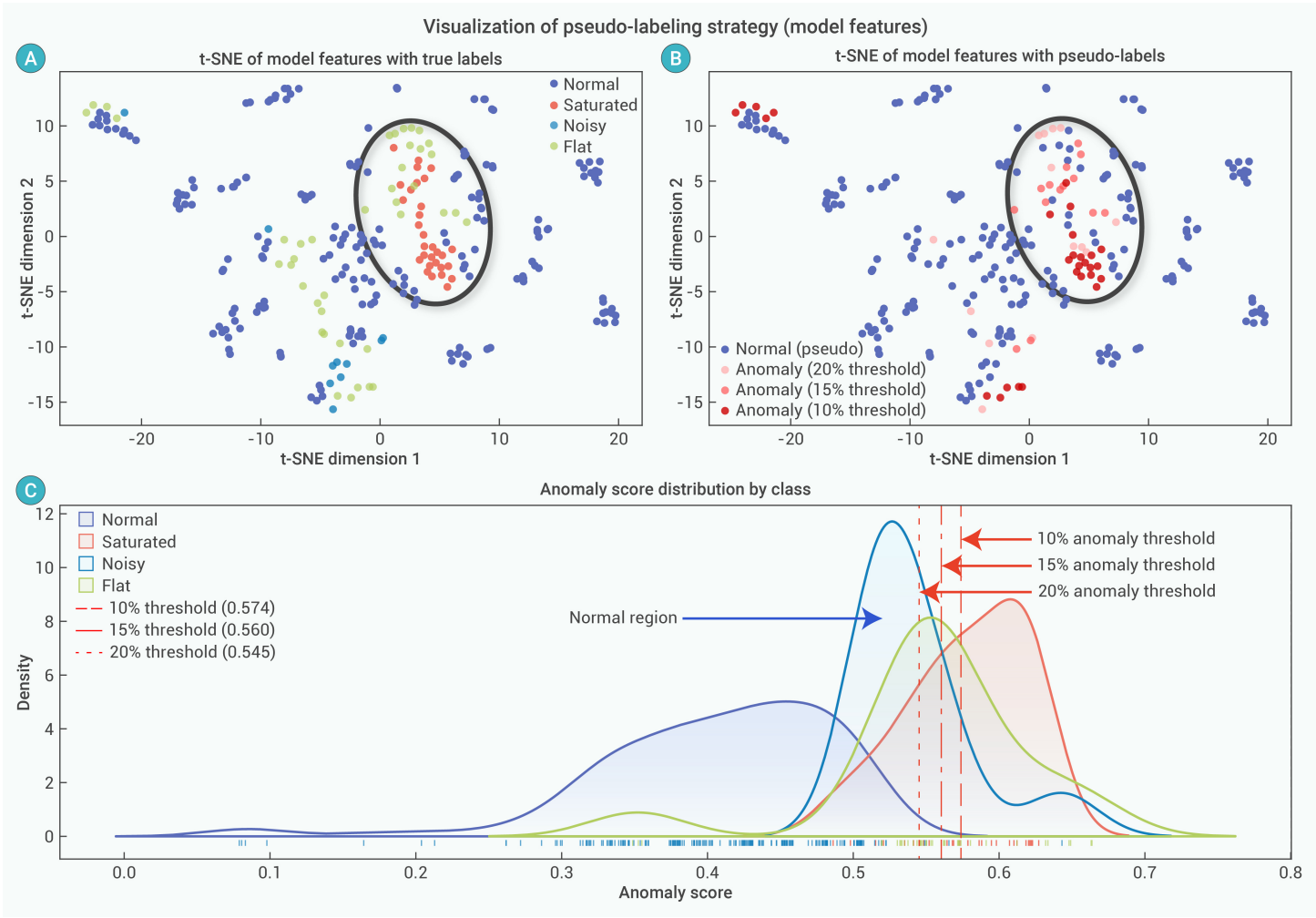


Figure 6. Visualization of pseudo-label generation mechanism (A) t-SNE visualization of model features colored by true labels, highlighting the inherent overlap between classes. (B) t-SNE visualization colored by generated pseudo-labels, illustrating the "easy-to-difficult" curriculum: the 10% threshold identifies the core Saturated cluster, while the 15% and 20% thresholds progressively expand to include boundary Noisy and Flat samples. (C) Anomaly score distributions for each true class. The vertical lines indicate the dynamic thresholds corresponding to the 10%, 15%, and 20% anomaly ratios, demonstrating how the progressive strategy captures samples from high certainty to high ambiguity.

samples (red squares) residing at the interface between them. This fundamental separability validates that our physics-guided learning paradigm has successfully sculpted a feature representation that is inherently and robustly discriminative.

Building on this, the 2D t-SNE visualization (Figure 7A) allows for a more detailed examination of the inference mechanism. The background contour visually represents the Mahalanobis distance-based anomaly score, creating a clear decision surface. The red dashed line delineates the explicit decision boundary, determined solely from the training set's normal data distribution. This panel provides direct visual evidence of how the diagnostic classifier operates, partitioning the space into a high-density normal region and a sparse anomalous region.

The quantitative performance is detailed in the confusion matrix (Figure 7C). The framework achieves a state-of-the-art F1-score of 99.0% and an accuracy of 99.6%. The model exhibits strong precision at 98.0% and perfect sensitivity with a recall of 100.0%, successfully identifying all 48 out of 48 anomalous spectra. This outstanding balance is achieved under the strict, reproducible 5-fold cross-validation protocol and using the empirically-justified 99.0% inference threshold (as detailed in "representation learning" Section), highlighting the practical robustness of the method.

A detailed analysis of the misclassified samples provides deeper insight into the model's behavior at the decision boundary. As shown in the confusion matrix (Figure 7C), the model achieves perfect recall, resulting in zero false negatives. This demonstrates that the PAQC framework, guided by the 99.0% threshold, successfully identifies all true anomalies, which is the primary goal of a quality control system.

There is only one false positive—a Normal spectrum incorrectly flagged as an anomaly. As shown in its spectral plot (Figure 7C, right) and its position in the t-SNE view (Figure 7A, red square), this is an "atypical" normal sample whose features lie just outside the conservative decision boundary. This single error, occurring at the class interface, indicates the model's high sensitivity to even subtle deviations from the core normal pattern. This near-perfect result underscores that the model does not suffer from systemic failure but rather accurately and robustly separates the normal and anomalous clusters.

In summary, the combination of a fundamentally discriminative feature space, sculpted by our hybrid-intelligence approach, and an effective, rigorously-defined distance-based classifier results in state-of-the-art performance. These results provide a comprehensive and honest validation of the PAQC paradigm, demonstrating its ability to translate a synergistic fusion of explicit physical knowledge and implicit deep learning into a reliable and automated solution for engineering quality control.

Comparison with state-of-the-art methods. We compare our PAQC framework against three widely-used anomaly detection baselines: Isolation Forest (IF), One-Class SVM (OC-SVM), Deep Support Vector Data Description (Deep SVDD), and Variational Autoencoder (VAE).[32] Performance is evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-score. All experiments were conducted using PyTorch, and detailed hyperparameter settings for all models are provided in the Supplementary Information, Section S3.2, to ensure reproducibility.

Before presenting the comparative results, it is crucial to re-emphasize that all performance metrics were obtained using the rigorous 5-fold cross-vali-

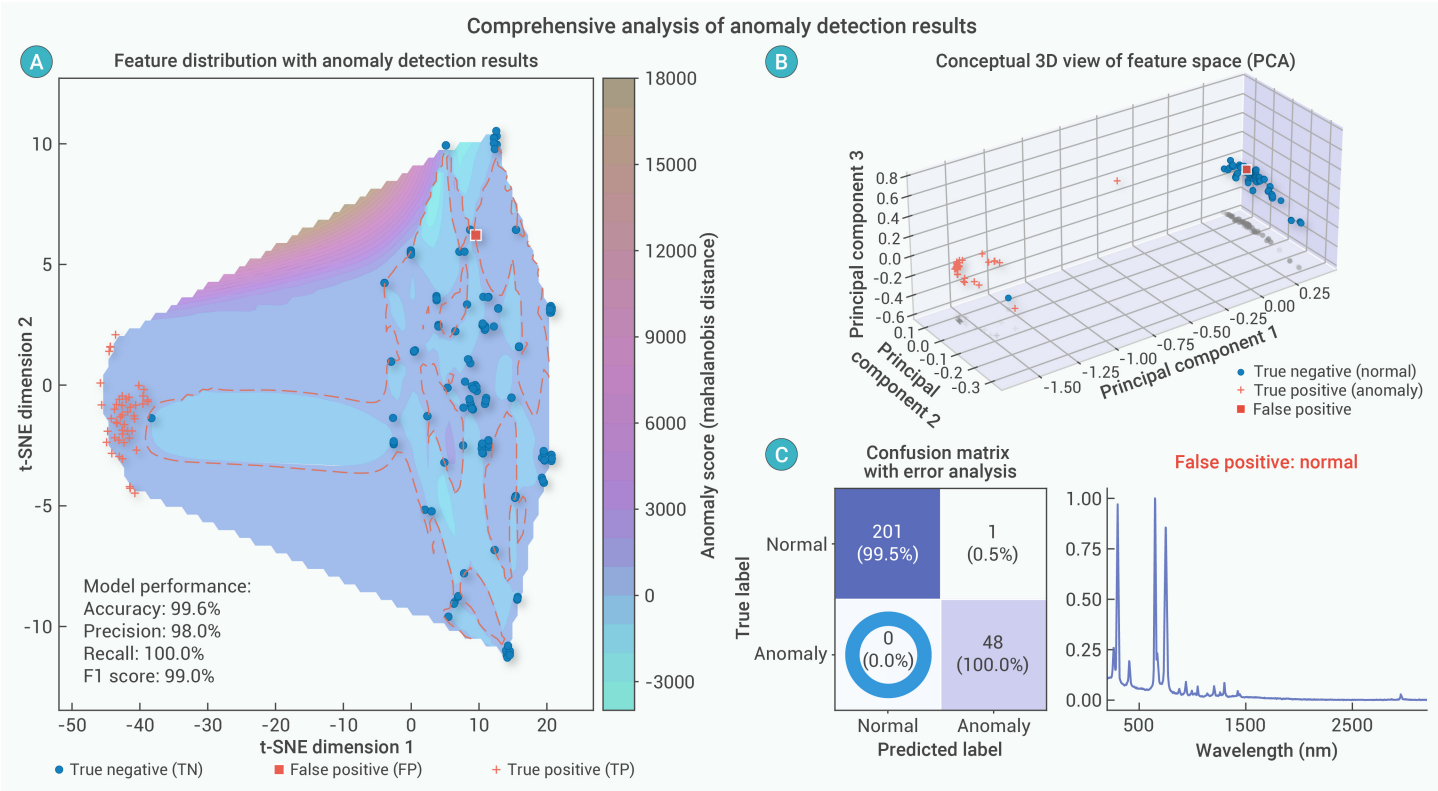


Figure 7. Visualization of the final anomaly detection results and the real-time diagnostic mechanism (A) 2D t-SNE visualization (Left) overlaid with the Mahalanobis distance-based anomaly score contour. The red dashed line delineates the explicit decision boundary determined by the 99.0th percentile of the training set, partitioning the space into a high-density normal region and a sparse anomalous region. (B) Conceptual 3D PCA view of the learned feature space (Top Right), demonstrating the clear linear separability between Normal (blue) and Anomalous (red) clusters. (C) Confusion matrix with error analysis (Bottom Right). The model achieves 100% recall. The single False Positive (Normal sample misclassified as Anomaly) is analyzed, showing it lies just outside the conservative decision.

dation (CV) protocol detailed in "Dataset" Section. All results, including for baselines, are reported as the mean ± standard deviation (*mean* ± *std*). This protocol ensures a rigorous and unbiased evaluation, guaranteeing that the following results are a genuine validation of each model's ability to generalize to previously unobserved data.

As shown in the baseline comparison section of Table 1, our PAQC framework significantly outperforms all baseline paradigms. While traditional methods like Isolation Forest (F1: 63.41% ± 1.55%) and One-Class SVM (F1: 75.25% ± 2.24%) struggle to model the complex spectral data, the deep learning approaches provide much stronger benchmarks.

The deep representation-based Deep SVDD, after extensive optimization, achieves a highly competitive F1-score of 93.73% ± 1.67%. To evaluate the generative paradigm, we implemented a deep VAE, which achieves an even higher F1-score of 96.85% ± 0.58%. However, despite its high overall performance, the VAE is limited by its recall (96.25% ± 2.04%). This indicates that while effective, pure generative models still struggle to detect "hard samples"—subtle, boundary-case anomalies that lack sufficient reconstruction error to be flagged without explicit guidance.

In contrast, our PAQC method, representing a hybrid-intelligence paradigm, achieves a superior F1-score of 98.98% ± 0.67% while maintaining a perfect 100.00% ± 0.00% recall. This performance gap, particularly in recall, is attributed to PAQC's synergistic design: unlike pure data-driven models (Deep SVDD and VAE), PAQC's physics-guided pseudo-labeling and contrastive learning actively leverage domain knowledge to identify and capture even the most elusive anomalies that other methods miss. A detailed analysis of all baseline models is available in the Supplementary Information, Section S3.

Ablation study: The role of the progressive strategy. The pivotal role of our progressive pseudo-labeling strategy is quantitatively assessed in the ablation study section of Table 1. A key finding is the model's high sensitivity to pseudo-label quality. This is a direct consequence of our semi-supervised contrastive learning design, where pseudo-labels are not merely for initialization but actively define the positive and negative pairs that guide the feature space optimization.

The results from the 5-fold cross-validation, presented in Table 1, reveal a critical insight: all tested strategies consistently achieve a perfect 100.00% ± 0.00% recall. The primary differentiator among the strategies, therefore, becomes precision, which reflects the quality of the learned decision boundary.

The fixed-proportion strategies clearly illustrate the challenge of selecting an optimal, fixed pseudo-label ratio, especially when the true anomaly ratio is unknown (assumed to be ~15% here). The 15% (Fixed) strategy, which is closest to the true anomaly ratio, yields the best F1-score (97.37% ± 1.04%) among all fixed approaches. The 10% (Fixed) strategy is too "conservative." By under-sampling the anomaly class, it fails to provide the model with enough diverse examples, resulting in a suboptimal F1-score (96.93% ± 1.15%). The 20% (Fixed) strategy is too "aggressive." By labeling 20% of samples as anomalous, it inevitably introduces "label noise"—i.e., mislabeling difficult normal samples as anomalies. This noise corrupts the feature space, leading to the lowest precision and the worst F1-score (96.36% ± 1.52%) among the fixed strategies.

Our progressive strategy (10%→15%→20%) masterfully resolves this trade-off by adopting an "easy-to-difficult" curriculum. It combines the strengths of all fixed strategies:Initial Stage (10%): It starts "conservatively," using only high-purity, unambiguous anomalies to establish a stable and robust feature space, avoiding the "noise" of the 20% strategy. Middle Stage (15%): It expands to the "optimal" 15% ratio, learning from a well-balanced set of samples. Final Stage (20%): Once the feature space is stable, it "aggressively" introduces the most challenging boundary cases, refining its discriminative power without succumbing to the noise that plagued the fixed 20% strategy from the start. Consequently, the progressive strategy achieves the highest precision (98.00% ± 1.31%) by most effectively mitigating label noise, resulting in the highest overall F1-score (98.98% ± 0.67%).

DISCUSSION

The experimental results comprehensively validate the state-of-the-art performance of the PAQC framework. More importantly, they provide a compelling solution to the dilemma between physics-based and data-driven

Table 1. Comprehensive performance evaluation: Comparison with state-of-the-art methods and ablation study of pseudo-labeling strategies.

Algorithm / Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Comparison with Baselines				
Isolation Forest	51.42 ± 2.03	47.34 ± 1.78	96.19 ± 2.65	63.41 ± 1.55
One-Class SVM	74.90 ± 2.81	66.40 ± 3.25	86.95 ± 1.81	75.25 ± 2.24
Deep SVDD	94.77 ± 1.33	98.12 ± 2.89	89.94 ± 4.25	93.73 ± 1.67
VAE	97.27 ± 0.41	97.50 ± 1.25	96.25 ± 2.04	96.85 ± 0.58
Ablation: Pseudo-label Strategies				
10% (Fixed)	97.21 ± 1.18	94.05 ± 2.15	100.00 ± 0.00	96.93 ± 1.15
15% (Fixed)	97.54 ± 1.05	94.88 ± 1.95	100.00 ± 0.00	97.37 ± 1.04
20% (Fixed)	96.73 ± 1.42	92.98 ± 2.81	100.00 ± 0.00	96.36 ± 1.52
Proposed Framework				
PAQC (Ours)	**99.11 ± 0.57**	**98.00 ± 1.31**	**100.00 ± 0.00**	**98.98 ± 0.67**

models for engineering signal analysis. PAQC successfully demonstrates that by synergizing physics-informed metrics with a powerful deep learning network, it is possible to achieve a system that is both robustly high-performing and grounded in domain expertise. The physics-based guidance provides reliability against noisy, unlabeled data, while the deep network provides the flexibility to learn complex patterns that defy simple physical modeling.

A key methodological insight from our ablation study is the importance of the knowledge fusion strategy. The success of the progressive pseudo-labeling curriculum highlights that for hybrid-intelligence systems, the mechanism for translating and feeding physics-based insights to the deep learning model must be a co-designed, integral component, not a simple preprocessing step. This approach of intelligently managing the uncertainty from physics-informed sources is critical for robust performance in real-world engineering applications.

Furthermore, the PAQC framework demonstrates significant potential for scalability across heterogeneous instrumentation (e.g., transferring from handheld to benchtop devices). This cross-instrument adaptability is underpinned by three pillars: (1) Universality of Physics-Informed Metrics: The fundamental definitions of spectral artifacts (e.g., saturation plateaus, baseline drift) are consistent across instrument types. Adapting to a new device requires only a simple, one-time calibration of the physical thresholds using standard reference samples. (2) Structural Resilience of the Autoencoder: The Peak-Attention Autoencoder is designed to be resolution-agnostic. Differences in spectral resolution or ranges between instruments can be bridged via standard spectral resampling (interpolation) to match the model's input dimension. Crucially, the peak-attention mechanism focuses on the topological structure of Raman peaks rather than their absolute pixel coordinates, ensuring robust feature extraction even under varying instrumental resolutions. (3) Efficient Transfer Learning: The framework's modular design enables rapid adaptation. The heavy feature extractor (Autoencoder) can be frozen, while the lightweight contrastive projection head—which contains significantly fewer parameters—can be efficiently fine-tuned. This allows the model to adapt to a new instrument's specific noise profile using only a small amount of data and minimal training time, avoiding the need for expensive full-model retraining.

The primary limitation of this work, which also points to a clear direction for future research, lies in the domain-specific nature of the explicit knowledge component. While the hybrid-intelligence paradigm of PAQC is highly generalizable, the specific physical indicators (SI, SBI, NI) are tailored for Raman spectroscopy. Applying this paradigm to other engineering domains, such as vibration analysis for mechanical fault diagnosis or signal integrity checking in communications, would necessitate the development of a new, domain-specific set of physics-informed indicators. This highlights a promising avenue for future work: collaborating with experts in various engineering fields to co-design new explicit knowledge modules and apply the PAQC blueprint to a wider range of knowledge-intensive quality control challenges.

CONCLUSION

In this work, we introduced and validated PAQC, a novel hybrid-intelligence paradigm for automated quality control in engineering applications utilizing Raman spectroscopy. Our framework addresses the fundamental dilemma between interpretable but simplistic physics-based models and powerful but knowledge-agnostic data-driven models. The core contribution is a framework that successfully fuses physics-informed metrics with a self-supervised deep learning model.

Through a synergistic design, this work demonstrated how physics-based insights can provide robust supervisory signals to effectively guide a deep contrastive learning network. Extensive experiments validated this approach, with the PAQC framework achieving a state-of-the-art F1-score of 98.98%. This work demonstrates that by strategically combining domain-specific physical insights with advanced deep learning techniques, it is possible to create a robust and automated solution for engineering quality control. The core principle of PAQC—synergizing domain knowledge with advanced representation learning—is highly generalizable. It offers a promising blueprint for developing robust, unsupervised decision-support systems for other analytical instrumentation and knowledge-intensive engineering applications where data quality is paramount and labeled anomalies are scarce.

REFERENCES

1. Hutsebaut D., Vandenabeele P. and Moens L. (2005). Evaluation of an accurate calibration and spectral standardization procedure for Raman spectroscopy. *Analyst* **130**:1204–1214. DOI:10.1039/B503624K
2. Nassif A.B., Talib M.A., Nasir Q., et al. (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access* **9**:78658–78700. DOI:10.1109/ACCESS.2021.3083060
3. Nardecchia A., Motto-Ros V. and Duponchel L. (2021). Saturated signals in spectroscopic imaging: Why and how should we deal with this regularly observed phenomenon. *Anal. Chim. Acta* **1157**:338389. DOI:10.1016/j.aca.2021.338389
4. Vulchi R.T., Morgunov V., Junjuri R., et al. (2024). Artifacts and anomalies in raman spectroscopy: A review on origins and correction procedures. *Molecules* **29**:4748. DOI:10.3390/molecules29194748
5. Wei D., Chen S. and Liu Q. (2015). Review of fluorescence suppression techniques in raman spectroscopy. *Appl. Spectrosc. Rev.* **50**:387–406. DOI:10.1080/05704928.2014.999936
6. He S., Zhang W., Liu L., et al. (2014). Baseline correction for Raman spectra using an improved asymmetric least squares method. *Anal. Methods* **6**:4402–4407. DOI:10.1039/C4AY00068D
7. Xu Y., Du P., Senger R., et al. (2021). ISREA: An efficient peak-preserving baseline correction algorithm for Raman spectra. *Appl. Spectrosc.* **75**:34–45. DOI:10.1177/0003702820955245
8. Guo S., Bocklitz T. and Popp J. (2016). Optimization of Raman-spectrum baseline correction in biological application. *Analyst* **141**:2396–2404. DOI:10.1039/C6AN00041J
9. Smulko J.M., Dingari N.C., Soares J.S., et al. (2014). Anatomy of noise in quantitative biological Raman spectroscopy. *Bioanalysis* **6**:411–421. DOI:10.4155/bio.13.337
10. Moester M.J.B., Ariese F. and de Boer J.F. (2015). Optimized signal-to-noise ratio with shot noise limited detection in stimulated Raman scattering microscopy. *J. Eur. Opt.*

Soc.-Rapid Publ. **10**:15022. DOI:10.2971/jeos.2015.15022
11. Ranjan R., Costa G., Ferrara M.A., et al. (2022). Noises investigations and image denoising in femtosecond stimulated Raman scattering microscopy. *J. Biophotonics* **15**:e202100379. DOI:10.1002/jbio.202100379
12. Pérez-Jiménez A.I., Lyu D., Lu Z., et al. (2020). Surface-enhanced Raman spectroscopy: Benefits, trade-offs and future developments. *Chem. Sci.* **11**:4563–4577. DOI:10.1039/D0SC00809E
13. Fan X., Zeng Y., Zhi Y.-L., et al. (2021). Signal-to-noise ratio enhancement for Raman spectra based on optimized Raman spectrometer and convolutional denoising autoencoder. *J. Raman Spectrosc.* **52**:890–900. DOI:10.1002/jrs.6065
14. Smulko J., Wróbel M.S. and Barman I. (2015). Noise in biological Raman spectroscopy. *Proc. ICNF* **2015**:1–6. DOI:10.1109/ICNF.2015.7288562
15. Fang S., Wu S., Chen Z., et al. (2024). Recent progress and applications of Raman spectrum denoising algorithms in chemical and biological analyses: A review. *TrAC Trends Anal. Chem.* **172**:117578. DOI:10.1016/j.trac.2024.117578
16. Stöckel S., Kirchhoff J., Neugebauer U., et al. (2016). The application of Raman spectroscopy for the detection and identification of microorganisms. *J. Raman Spectrosc.* **47**:89–109. DOI:10.1002/jrs.4844
17. Liu M., Wang T., Zhang Q., et al. (2024). An outlier removal method based on PCA-DBSCAN for blood-SERS data analysis. *Anal. Methods* **16**:846–855. DOI:10.1039/D3AY02037A
18. Li C., Guo L., Gao H., et al. (2021). Similarity-measured isolation forest: Anomaly detection method for machine monitoring data. *IEEE Trans. Instrum. Meas.* **70**:1–12. DOI:10.1109/TIM.2021.3062684
19. Strani L., Cocchi M., Tanzilli D., et al. (2025). One class classification (class modelling): State of the art and perspectives. *TrAC Trends Anal. Chem.* **183**:118117. DOI:10.1016/j.trac.2024.118117
20. Kelis Cardoso V.G. and Poppi R.J. (2021). Cleaner and faster method to detect adulteration in cassava starch using Raman spectroscopy and one-class support vector machine. *Food Control* **125**:107917. DOI:10.1016/j.foodcont.2021.107917
21. Zhang L., Huang X., Li P., et al. (2017). Multivariate adulteration detection for sesame oil. *Chemometr. Intell. Lab. Syst.* **161**:147–150. DOI:10.1016/j.chemolab.2016.11.009
22. Song S., Liu Z., Huang M., et al. (2020). Detection of fish bones in fillets by Raman hyperspectral imaging technology. *J. Food Eng.* **272**:109808. DOI:10.1016/j.jfoodeng.2019.109808
23. Feng Q., Zhang Y., Li C., et al. (2017). Anomaly detection of spectrum in wireless communication via deep auto-encoders. *J. Supercomput.* **73**:3161–3178. DOI:10.1007/s11227-017-2017-7
24. Mhamdi L., McLernon D., El-moussa F., et al. (2020). A deep learning approach combining autoencoder with one-class SVM for DDoS attack detection in SDNs. *IEEE Int. Conf. Commun.* **2020**:1–6. DOI:10.1109/ComNet47917.2020.9306073
25. Principi E., Rossetti D., Squartini S., et al. (2019). Unsupervised electric motor fault detection by using deep autoencoders. *IEEE/CAA J. Autom. Sinica* **6**:441–451. DOI:10.1109/JAS.2019.1911393
26. Zhou Y., Liang X., Zhang W., et al. (2021). VAE-based Deep SVDD for anomaly detection. *Neurocomputing* **453**:131–140. DOI:10.1016/j.neucom.2021.04.089
27. Zhang Z. and Deng X. (2021). Anomaly detection using improved deep SVDD model with data structure preservation. *Pattern Recognit. Lett.* **148**:1–6. DOI:10.1016/j.patrec.2021.04.020
28. Wei J., Wang Q., Zhang G., et al. (2025). Domain knowledge guided pseudo-label generation framework for semi-supervised domain generalization fault diagnosis. *Adv. Eng. Inf.* **67**:103540. DOI:10.1016/j.aei.2025.103540
29. Sun Y., Tao H. and Stojanovic V. (2025). Pseudo-label guided dual classifier domain adversarial network for unsupervised cross-domain fault diagnosis with small samples. *Adv. Eng. Inf.* **64**:102986. DOI:10.1016/j.aei.2024.102986
30. Li Y.-F., He W.-Y., Ren W.-X., et al. (2025). Moving load induced dynamic response analysis of bridge based on physics-informed neural network. *Adv. Eng. Inf.* **65**:103215. DOI:10.1016/j.aei.2025.103215
31. Zhu Y., Li S., Lang X., et al. (2025). Physics-informed CGAN and multi-scale attention CNN for pipeline leakage diagnosis under imbalanced data. *Adv. Eng. Inf.* **66**:103471. DOI:10.1016/j.aei.2025.103471
32. Kingma D.P. and Welling M. (2022). Auto-Encoding variational bayes. *arXiv preprint.* DOI:10.48550/arXiv.1312.6114

FUNDING AND ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. T2322020, 62371329,22174133, 32202540, 51832005 and 1210042018), and the National Key R&D Plan of China (No.2021YFC3340400, No. 2024YFF1206500 and No. 2024YFF1206504) the Tianjin Young Scientific and Technological Talents (Grant No. QN20230107), Natural Science Foundation of Tianjin (Grant No. 24JCQNJC00810),and the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02125). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Xin Chen: Conceptualization, Writing – review & editing, Writing – original draft, Methodology, Investigation. Zilong Wang: Writing – review & editing, Supervision, Methodology. Yuting Li: Writing – review & editing, Resources. Liang Li: Investigation, Writing - Review & Editing. Biao Sun: Conceptualization, Supervision, Funding acquisition. Pei Liang: Writing–review & editing, Supervision, Funding acquisition, Conceptualization, Project Administration. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Biao Sun is an Academic Editor of The Innovation journals and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.

DATA AND CODE AVAILABILITY

The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-inform.2026.100028.