IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Instrumentation Mesure Métrologie
Vol. 25, No. 5, October, 2026, pp. 271-281
Journal homepage: http://iieta.org/journals/i2m

A Fast Linear Spectroscopic Retrieval Framework for Exhaled Acetone Using HITRAN Cross Section Data for Non-Invasive Diabetes Screening

Zeyad Al-Ibadi*, M. H. Asmaa, Buraq Kareem Kadhim

Department of Laser Physics, College of Science for Women, University of Babylon, Babylon 51001, Iraq

Corresponding Author Email: wsci.ziead.khalaf@uobabylon.edu.iq

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/i2m.250503

Received: 8 July 2026
Revised: 16 September 2026
Accepted: 26 September 2026
Available online: 31 October 2026

Keywords:
acetone breath analysis, HITRAN, linear regression, diabetes screening, point-of-care sensing, spectroscopic simulation

ABSTRACT

Exhaled acetone is a promising non-invasive biomarker for type 1 diabetes and ketotic states. This study presents a physically interpretable linear retrieval framework that estimates acetone concentration from the peak absorbance of spectra simulated from HITRAN2020 cross-section data using the Beer-Lambert law. A dataset of 2,000 breath-like spectra was generated with acetone concentrations of 0.1–20 ppm, temperatures of 20–40 °C, pressures of 0.95–1.05 atm, and a signal-to-noise ratio (SNR) of 20–50 dB. The proposed linear regression model achieved a coefficient of determination (R^2) = 0.979, mean absolute error (MAE) = 0.57 ppm, and root mean square error (RMSE) = 0.82 ppm on the independent test set, while requiring an average inference time of approximately 9 µs per sample. Statistical validation included 95% prediction intervals, residual diagnostics, sensitivity analysis with respect to temperature, pressure, and SNR, and receiver operating characteristic analysis at the assumed 2 ppm threshold (area under the ROC curve, ROC-AUC ≈ 0.99). This represents a simulation-based assessment of discrimination rather than validated clinical diagnostic performance. In addition, the framework was evaluated under a realistic noise model, demonstrating robust retrieval performance while supporting comprehensive statistical and metrological characterization. These results indicate that the proposed framework provides an efficient and transparent basis for quantitative breath-acetone retrieval. Experimental validation using real exhaled breath samples remains essential before clinical application.

1. INTRODUCTION

Continuous research interest has been focused on analysing exhaled breath as a diagnostic tool since the early 2000s [1, 2]. There are hundreds of volatile organic chemicals (VOCs) in human breath, with some of these receiving particular interest due to their connection with ketosis conditions and type 1 diabetes, such as acetone [3, 4]. Normal levels of acetone in breath for healthy individuals are between 0.2 and 1 ppm. A high concentration of >2 ppm is generally regarded as an indicator of insulin deficiency or increased production of ketone bodies [5-7].

Tunable diode laser absorption spectroscopy (TDLAS) and cavity-enhanced techniques have demonstrated tremendous potential in the field of real-time breath monitoring [8, 9]. However, these methods have their own limitations, requiring precise calibration of the hardware and signal processing operations generally based on non-linear least-squares fitting and/or neural networks. Such methods are typically implemented through the use of commercial packages, such as the MATLAB Deep Learning Toolbox [10, 11].

Several groups have used simulated absorption spectra of breath gases such as acetone from the HITRAN (High-Resolution Transmission Molecular Absorption Database) [8, 12].

In particular, the 3 µm region of HITRAN2020 is well suited to laser-based sensing of acetone [13]. However, published literature on this topic is concentrated on instrument design or on providing the raw spectra as input to black-box machine learning models, without a full report of the statistical framework of the results, such as prediction intervals, normality of the residuals, and sensitivity to temperature, pressure, and signal-to-noise ratio (SNR).

In parallel with recent developments in machine-learning-based spectroscopy, quantitative absorption spectroscopy has continued to rely on physically interpretable retrieval methods derived from the Beer-Lambert law. Linear inversion and single-feature regression approaches remain widely used for quantitative gas analysis because they establish a direct relationship between absorbance and concentration while maintaining low computational complexity, transparent calibration, and straightforward uncertainty analysis. The present work follows this established retrieval philosophy by combining HITRAN cross-section data with Beer-Lambert modelling and a linear calibration framework for rapid acetone concentration estimation [14-19].

Combining simple, linear, interpretable HITRAN-based acetone retrieval with confidence intervals, residual analysis, receiver operating characteristic (ROC) analysis, and temperature/pressure/SNR sensitivity at the 2 ppm threshold

used in the simulation has received limited attention in the authors' knowledge. Computational time and hardware independence are also rarely considered for point-of-care (POC) translation [20-22].

This work presents four contributions. First, we propose a linear retrieval framework for acetone based on peak absorbance from HITRAN data, supported by a complete statistical pipeline. Second, to our knowledge, no earlier acetone study has combined a 95% prediction interval, a ROC curve and its area under the curve (AUC) and a three-factor sensitivity analysis (SNR, temperature, pressure) in a single study. Third, the implementation relies only on core MATLAB functions without requiring additional toolboxes, making the proposed workflow readily reproducible across standard MATLAB installations. Finally, performance is also reported against a clinical screening threshold of 2 ppm using cumulative error curves, which ties the numerical accuracy directly to a clinical decision boundary.

2. MATERIALS AND METHODS

2.1 Proposed measurement architecture for future experimental implementation

Although the present work is entirely simulation-based, a conceptual measurement architecture is proposed to facilitate future experimental implementation and to place the present retrieval framework within a complete metrological measurement chain. As shown in Figure 1, exhaled breath is collected through a sampling interface, passed through a moisture-removal stage to reduce water-vapor interference, and introduced into an optical gas cell illuminated by a mid-infrared laser operating near the acetone absorption band (~2,970 cm^{-1}). The transmitted signal is detected by an infrared detector, digitized, and processed to extract the peak absorbance, from which acetone concentration is estimated using the linear calibration model developed in this study. Prior to routine operation, the system would be calibrated using certified acetone gas mixtures, with periodic recalibration performed to maintain calibration validity and long-term stability [23-26].

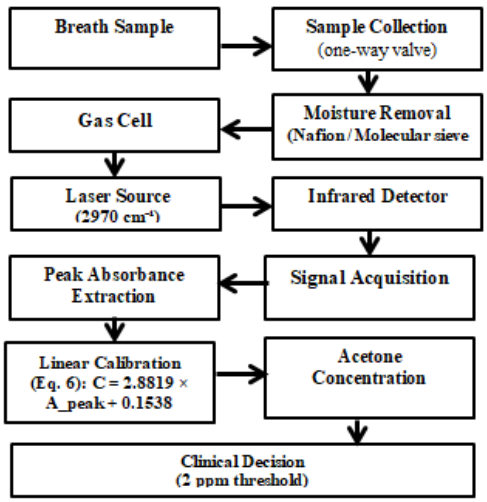


Figure 1. Proposed measurement architecture for future experimental implementation of the acetone retrieval framework

2.2 HITRAN data acquisition

Cross-sections for acetone (C_3H_6O, molecular ID 55) were downloaded from the HITRAN2020 online interface (https://hitran.org/xsc/) [8]. The chosen file covers 2,615–3,300 cm^{-1}, roughly the 3.0–3.8 μm window. The recording conditions were temperature = 296 K, pressure = 759 Torr, with 136,400 data points sampled at 0.005 cm^{-1}. Two-column plain text format was used, with wavenumber in cm^{-1} and cross-section in $cm^2 \cdot molecule^{-1}$. No re-reduction was applied at any stage, so that the original spectral resolution was preserved.

2.3 Generation of spectra based on HITRAN cross-sections and the Beer-Lambert law

A dataset comprising 2,000 simulated breath-related spectra was generated by implementing the Beer-Lambert relationship while incorporating clinically plausible variations in the model parameters. The absorbance A at wavenumber v is expressed as Eq. (1).

$$A_{(\nu)} = \sigma(\nu) \cdot C \cdot L \tag{1}$$

where, $\sigma(\nu)$ is the cross section from the HITRAN database ($cm^2 \cdot molecule^{-1}$), C is the molecular number density ($molecules \cdot cm^{-3}$), and L is the optical path length, fixed at 0.1 cm to match small-sized gas cell geometries. The concentration was converted from ppm to $molecules \cdot cm^{-3}$ using the ideal gas law, Eq. (2).

$$C = \frac{[ppm] \times 10^{-6} \cdot P_{atm} \cdot 101325}{K_B \cdot T} \tag{2}$$

where, K_B = 1.380649 × 10^{-23} $J \cdot K^{-1}$, P_{atm} in atm, and T in Kelvin.

For each simulated spectrum, all four parameters were sampled independently from a uniform distribution.

The range of acetone concentrations varied from 0.1 ppm to 20 ppm, which corresponded to normal physiological and diabetic or ketotic levels of acetone. Given that the temperature range affects performance from ambient room temperature through exhaled breath temperatures, it was proposed that temperature be varied in the range of 20 to 40 °C. The pressure fluctuated between 0.95 atm and 1.05 atm, which is the normal change in pressure in the atmospheric environment. The SNR was selected in the range of 20–50 dB.

The spectra were then corrupted with additive Gaussian noise, with σ_{noise} set from each spectrum's peak amplitude and the chosen SNR as given by Eq. (3).

$$\sigma_{noise} = \frac{signal_{max}}{10^{SNR/20}} \tag{3}$$

where, $signal_{max}$ is that peak amplitude, so noise scales with signal strength rather than remaining fixed. The random number generator (RNG) seed was fixed at 42 using `rng(42)` for reproducibility.

The synthetic spectra were also assessed with an improved noise model that includes the original additive detector noise to better represent real measurements. The enhancements to the model were based on the common sources of uncertainty in infrared spectroscopic instrumentation, which included multiplicative intensity fluctuations, linear baseline drift, and

small shifts in the wavenumber axis, which were considered as a simpler model of the wavelength calibration uncertainty. The retrieval model was subsequently further tested with 20 paired repetitions under the same realization of concentration, temperature, pressure and SNR, the effect of the additional perturbations being quantified by comparing the results of the two retrievals. To assess the performance, the resulting performance measures are the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE), which are shown in Section 3.11.

2.4 Feature extraction and data split

For each noisy spectrum, the maximum value across the wavenumber axis was taken as the single feature for the retrieval, hereafter referred to as the peak absorbance. This feature was selected for two reasons: first, the Beer-Lambert law itself imposes a linear relationship between peak absorbance and concentration, so the selected feature is directly aligned with the underlying physics. Second, locating a maximum carries a negligible computational cost, which keeps the retrieval compatible with embedded hardware. The 2,000 spectra were then split randomly, with 1,400 assigned to training, 300 to validation, and 300 held out for testing (70/15/15).

2.5 Linear regression model

By using MATLAB's *polyfit* function, a first-degree polynomial was fitted to the training data, and the model takes the form of Eq. (4).

$$\hat{C} = a \cdot A_{peak} + b \tag{4}$$

where, $\hat{C}$ is the predicted concentration in ppm, A_{peak} is the peak absorbance, and a and b are the slope and intercept, respectively. Predictions on the validation and test sets were obtained through *polyval*. Neither iterative optimisation nor any external toolbox was required.

2.6 Performance metrics

Three standard metrics were used to assess model performance: R^2 (defined in Eq. (5)), MAE in ppm, and RMSE in ppm.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \tag{5}$$

2.7 Statistical analysis and additional diagnostics

To complement the statistical performance metrics, a first-order uncertainty propagation analysis was performed following the Guide to the Expression of Uncertainty in Measurement (GUM). Since acetone concentration is obtained by inversion of the Beer-Lambert relationship, the contributions of temperature, pressure, and optical path length uncertainties were propagated using a first-order Taylor expansion. The corresponding sensitivity coefficients were derived analytically from the governing equations, and the individual uncertainty components were combined using the root-sum-of-squares (RSS) method to obtain the combined standard uncertainty. Expanded uncertainty was then calculated using a coverage factor of $k = 2$ (approximately 95% confidence) [27, 28].

The prediction interval follows the standard formula for simple linear regression as Eq. (6).

$$se_{pred} = \sqrt{\text{MSE}\,(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}})} \tag{6}$$

where, MSE = SSE/(n−2) is the mean squared error in Eq. (6), SSE = $\Sigma(y_i - \hat{y}_i)^2$ is the sum of squared errors, $\bar{x}$ is the mean true concentration, and $S_{xx} = \Sigma(x_i - \bar{x})^2$ is its corrected sum of squares. The formula assumes normally distributed, homoscedastic residuals.

For n = 300 (298 degrees of freedom), the exact Student's t critical value is $t(0.975, 298) = 1.968$, within 0.5% of 1.96, so the normal approximation is appropriate [27].

Several diagnostic procedures were carried out for model validation. A 95% prediction interval for individual test samples was derived from the standard error of prediction.

Residual analysis comprised the inspection of residuals against fitted values, residuals in order of measurement, and a histogram overlaid with a normal density curve. A Q-Q plot with normal quantiles obtained through erfinv was generated for normality assessment. Autocorrelation of residuals was tested by means of the Durbin-Watson statistic.

Sensitivity analysis was performed by recomputing MAE while varying one parameter at a time: SNR (20–50 dB), temperature (20–40 °C), and pressure (0.95–1.05 atm).

Pearson correlation between concentration and five spectral features—peak, AUC, mean, median, and standard deviation—was also examined. The influence of optical path length was examined at 0.05, 0.1, 0.2, 0.5, and 1.0 cm.

For the assumed 2 ppm threshold, an ROC curve was generated by computing true and false positive rates over 50 decision values; the AUC was integrated using the trapezoidal rule.

The proposed method was also compared qualitatively with existing methods considered in the computational modelling of acetone breath analysis. All computations were performed on an AMD A10-5750M APU (2.50 GHz, 4 GB RAM) using MATLAB R2019a. Computations were executed in a single MATLAB session on a single CPU thread without explicit parallelization or multithreading. The average inference time was measured using MATLAB's *tic*/*toc* functions over 1,000 repeated evaluations.

3. RESULTS

3.1 Model performance

The linear regression model fitted to the peak absorbance gave consistent behaviour across all three subsets, as shown in Table 1.

Table 1. Model performance on training, validation, and test sets

Dataset	R^2	MAE (ppm)	RMSE (ppm)
Training	0.9819	0.5517	0.7947
Validation	0.9800	0.5451	0.8258
Test	0.9791	0.5665	0.8191

Note: Coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE).

R^2 stayed above 0.98 in all three subsets. In plain terms, a single feature—the peak absorbance—explains over 98% of the variance in acetone concentration.

Additionally, the gap between validation-set and test-set performance is negligible, within 0.55–0.57 ppm, with a test RMSE of 0.82 ppm, consistent with the absence of overfitting [9]. The model equation obtained from the training set is Eq. (7).

$$\hat{C}_{acetone} = 2.8819 \cdot A_{peak} + 0.1538 \quad (7)$$

3.2 Calibration curve and prediction interval

Figure 2 presents the test-set predicted concentrations plotted against the corresponding true values.

It displays the ideal identity ($y = x$) as a solid black line, the fitted regression as a dashed blue line, and the 95% prediction interval as a grey shaded area.

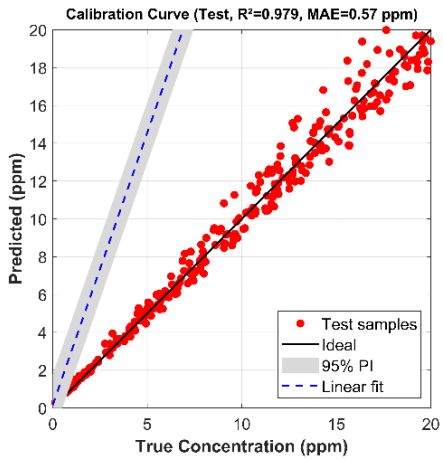


Figure 2. Calibration curve for the test set: predicted vs. true acetone concentration

The data points lie close to the identity line, indicating good agreement between the predicted and true concentrations. The prediction interval remains narrow, widening to approximately ±1.6 ppm at the extremes of the tested range. Such a narrow interval is itself indicative of the model's precision [7].

Model-based metrological indicators, derived from the simulated spectra and the fitted calibration model rather than from a physically characterized measurement instrument, were computed to characterize simulation-derived analytical performance. The sensitivity coefficient is the slope of Eq. (7), $S = 2.8819$ ppm per unit peak absorbance. The simulated limit of detection (LOD) and limit of quantification (LOQ) were computed using the standard 3.3σ and 10σ criteria, with σ taken as the test-set residual standard deviation (0.8193 ppm, reflecting the full tested SNR range of 20–50 dB), yielding LOD = 2.70 ppm, LOQ = 8.19 ppm. The blended LOD exceeds the 2 ppm clinical threshold across the tested SNR range (20–50 dB), indicating that near-threshold detection may require a higher SNR.

Under best-case conditions (SNR = 50 dB, Section 3.4), where MAE falls to 0.013 ppm, the corresponding LOD and LOQ are approximately 0.05 ppm and 0.16 ppm, respectively [29]. These two estimates are conceptually distinct: the LOD/LOQ values under the full SNR range are derived from the regression residual standard deviation, whereas the best-case values are inferred from the MAE achieved under a specific simulated SNR condition. The best-case scenario is illustrative of the simulation's operating envelope and should not be interpreted as an instrument-validated detection capability. Both are reported for transparency (Table 2).

The dynamic range, taken as the ratio of the upper tested concentration (20 ppm) to the LOQ, is approximately 2.4:1 under blended average conditions and approximately 125:1 under best-case conditions.

Table 2. Metrological performance indicators

Indicator	Blended (Full SNR Range)	Best Case (SNR = 50 dB)
Sensitivity (ppm/absorbance unit)	2.8819	2.8819
LOD (3.3σ)	2.70 ppm	0.05 ppm
LOQ (10σ)	8.19 ppm	0.16 ppm
Dynamic range (20 ppm/LOQ)	~2.4:1	~125:1

Note: Limit of detection (LOD), limit of quantification (LOQ), signal-to-noise ratio (SNR).

3.3 Residual analysis

The diagnostics for test-set residuals shown in Figure 3 are four plots. No funnel pattern was observed in the residuals plotted against the fitted values, indicating the absence of practically significant heteroscedasticity. Points in the middle region were close to the diagonal line, with only some minor deviations in the tails in a Q-Q plot. There was no trend when the residuals were plotted against the order of measurement, and the Durbin-Watson statistic was 2.085 (Table 3), close to the expected value in the absence of autocorrelation.

Table 3. Summary statistics of the test-set residuals

Statistic	Value
Mean residual (ppm)	−0.0439
Skewness	−0.8954
Excess kurtosis	1.9605
Durbin-Watson statistic	2.0854
95th percentile of |residual| (ppm)	1.7923

Residuals are defined as the difference between the true and predicted concentrations ($y_i - \hat{y}_i$); therefore, a negative mean residual indicates a slight systematic over-prediction. The mean residual was −0.044 ppm, with a standard deviation of 0.819 ppm. In the residual histogram, the plotted distribution was nearly identical to the overlaid normal density curve.

The skewness was −0.895 and excess kurtosis was 1.96, meaning that the distribution was slightly left-skewed and had somewhat fatter tails than a Gaussian distribution. The absolute residuals between data and model were approximately normally distributed with a mean of zero; the 95th percentile of |residual| was 1.79 ppm [6].

3.4 Sensitivity to environmental variations

MAE exhibited a strong dependence on SNR but only a weak response to temperature or pressure. As SNR rose from 20 to 50 dB, MAE decreased from 0.300 ppm to 0.013 ppm, passing through 0.099 ppm at 30 dB and 0.040 ppm at 40 dB (Figure 4 and Table 4).

Temperature variations over 20–40 °C produced only marginal changes in MAE (0.038–0.043 ppm), and pressure variations over 0.95–1.05 atm yielded an even smaller swing

(0.0366–0.0389 ppm). The peak-absorbance linear retrieval may therefore be regarded as robust to the environmental variations considered in this study [10].

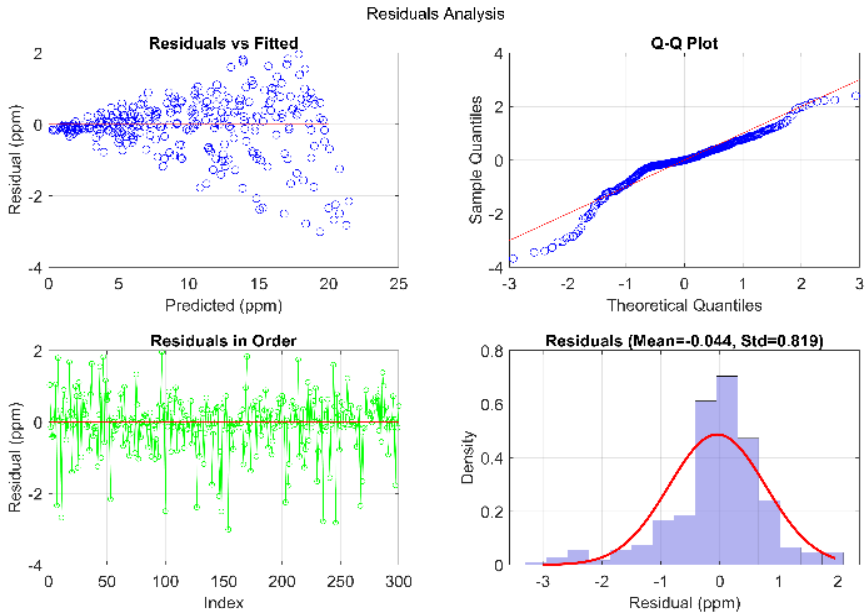


Figure 3. Residual diagnostic plots for the linear regression model on the test set

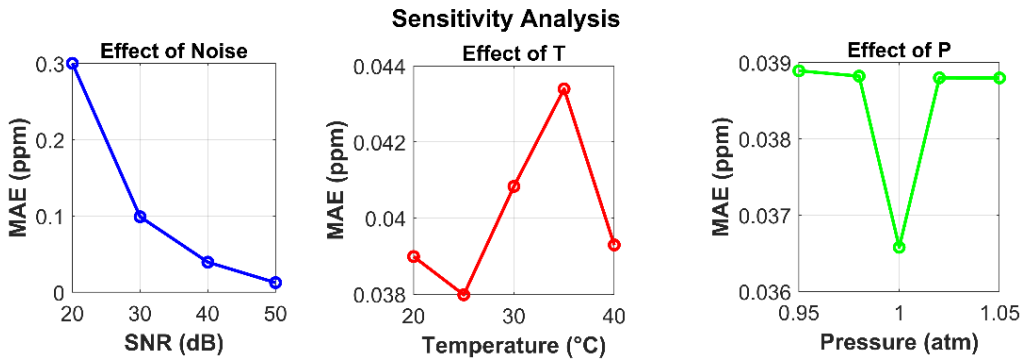


Figure 4. MAE as a function of SNR, temperature, and pressure
Note: Signal-to-noise ratio (SNR), mean absolute error (MAE).

Table 4. MAE values for varying SNR, temperature, and pressure

Condition	Value	MAE (ppm)
SNR (dB)	20	0.3002
SNR (dB)	30	0.0990
SNR (dB)	40	0.0395
SNR (dB)	50	0.0126
Temperature (°C)	20	0.0390
Temperature (°C)	25	0.0380
Temperature (°C)	30	0.0408
Temperature (°C)	35	0.0434
Temperature (°C)	40	0.0393
Pressure (atm)	0.95	0.0389
Pressure (atm)	0.98	0.0388
Pressure (atm)	1.00	0.0366
Pressure (atm)	1.02	0.0388
Pressure (atm)	1.05	0.0388

Note: Signal-to-noise ratio (SNR), mean absolute error (MAE).

3.5 Uncertainty propagation analysis

A model-based uncertainty budget was established using first-order Taylor uncertainty propagation, derived from the simulation framework and assumed input uncertainties rather than from a physically characterized measurement instrument. The contributions of temperature, pressure, and optical path length uncertainties were quantified under the reference operating conditions (temperature = 25 °C, pressure = 1 atm, L = 0.1 cm). Table 5 summarizes the individual and combined uncertainty contributions.

The combined standard uncertainty was found to be substantially smaller than the statistical prediction interval reported in Section 3.2, indicating that environmental parameter uncertainties contribute only a minor fraction of the total prediction uncertainty under the assumptions adopted in this study [27, 28].

3.6 Feature correlation

Pearson correlation coefficients between acetone concentration and five candidate spectral features are reported in Figure 5 and Table 6. The highest correlation (r = 0.9976) was shared by two features, namely the AUC and the mean

absorbance. These were followed in close succession by the peak (0.9896) and the standard deviation (0.9893). The median exhibited a markedly lower correlation of 0.7893.

Table 5. Uncertainty budget for temperature, pressure, and path length

Source	Standard Uncertainty	Sensitivity Coefficient	Relative Contribution	Absolute Contribution at 2 ppm	Absolute Contribution at 10 ppm
Temperature (T)	±1 °C	$\partial(\ln \text{ppm})/\partial T = +1/T$	±0.34%	±0.007 ppm	±0.034 ppm
Pressure (P)*	±0.01 atm (±1%)	$\partial(\ln \text{ppm})/\partial P = -1/P$	±0.99–1.01%	±0.020 ppm	±0.099–0.101 ppm
Path length (L)*	±1% of L	$\partial(\ln \text{ppm})/\partial L = -1/L$	±0.99–1.01%	±0.020 ppm	±0.099–0.101 ppm
Combined standard uncertainty, u_c (root sum square (RSS))	—	—	±1.47%	±0.029 ppm	±0.147 ppm
Expanded uncertainty, $U = k \cdot u_c$ ($k = 2$, ≈ 95%)	—	—	±2.94%	±0.059 ppm	±0.293 ppm
Statistical 95% prediction intervals (Section 3.2)	—	—	—	—	up to ±1.6 ppm

Note : *Assumed standard uncertainty, not independently measured for this study. From Eqs. (1) and (2), C ∝ AT/(PL) at fixed absorbance, giving $\partial \ln C/\partial T = +1/T$, $\partial \ln C/\partial P = -1/P$, and $\partial \ln C/\partial L = -1/L$. Higher temperature lowers gas density at fixed pressure, raising the concentration needed for a given absorbance, while higher pressure or path length has the opposite effect.

Although the area and the mean carried marginally stronger correlations than the peak, the peak was selected as the working feature on grounds of computational economy, since its extraction requires only a single max operation and is therefore the most efficient choice for embedded POC implementation [21].

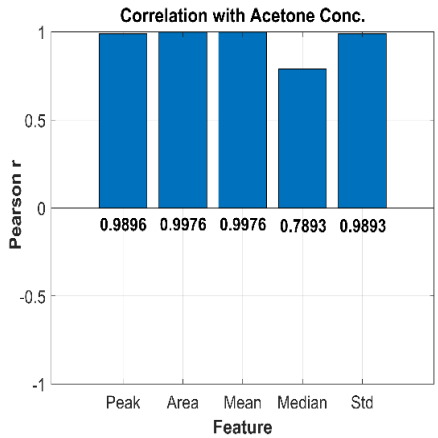


Figure 5. Pearson correlations between acetone concentration and five spectral features: peak absorbance, area, mean, median, and standard deviation

Table 6. Pearson correlation coefficients between spectral features and acetone concentration (test set)

Feature	Peak	Area	Mean	Median	Standard Deviation
Pearson r	0.9896	0.9976	0.9976	0.7893	0.9893

Table 7. Cumulative percentage of test samples within given absolute error thresholds

Error Threshold (ppm)	0.25	0.50	0.75	1.00	1.50	2.00
Cumulative Percentage (%)	39.7	58.3	73.0	84.3	91.0	96.0

3.7 Cumulative error distribution (clinical tolerance)

Figure 6 and Table 7 summarise the fraction of test samples whose absolute error fell below increasing tolerance thresholds. A fraction of 58.3% of predictions lay within ±0.5 ppm of the true value, 84.3% within ±1.0 ppm, and 96.0% within ±2.0 ppm. These percentages are consistent with the tolerance ranges considered clinically acceptable for diabetes screening in the published literature [13].

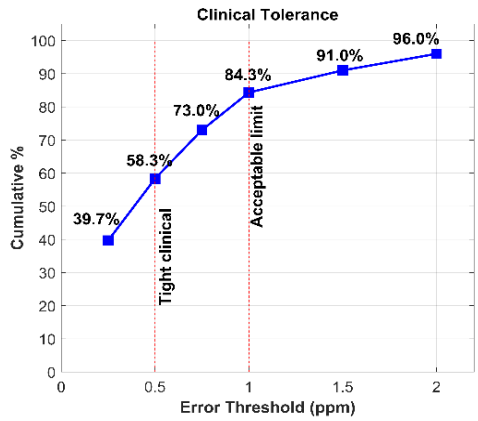


Figure 6. Cumulative percentage of test samples within given absolute-error thresholds (clinical tolerance)

3.8 Comparison of linear and quadratic models

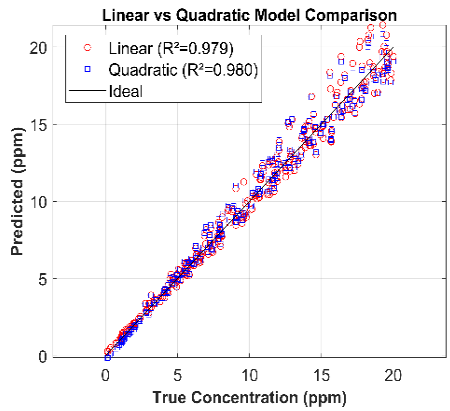


Figure 7. Comparison between linear ($R^2 = 0.9791$) and quadratic ($R^2 = 0.9793$) regression models on the test set

A quadratic polynomial was fitted to the same data for comparison (Figure 7). Both R^2 and MAE differed by less than 0.001 between the two fits. Within the clinical range examined here, the response is therefore linear in practice. The inclusion

of a quadratic term would offer no benefit while increasing the risk of overfitting [20].

3.9 Effect of path length

Figure 8 displays MAE as a function of optical path length, with SNR fixed at 40 dB, temperature at 25 °C, and pressure at 1 atm. The dependence is weak. MAE drops from about 0.0405 ppm at 0.2 cm to 0.0365 ppm at 1.0 cm—a marginal improvement that may not justify the increased cell length. At the selected path length of 0.1 cm, MAE was already 0.039 ppm, demonstrating that compact gas cells remain feasible for POC designs [21].

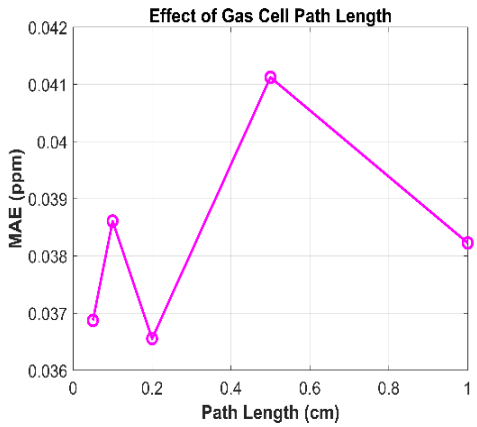


Figure 8. Effect of gas cell path length on MAE (SNR = 40 dB, temperature = 25 °C, pressure = 1 atm)
Note: Signal-to-noise ratio (SNR), mean absolute error (MAE).

To provide a physical interpretation of this behavior, the path-length dependence was re-examined analytically. According to the Beer-Lambert law, peak absorbance is directly proportional to the optical path length ($A \propto L$). Under the proportional noise model adopted throughout this study, detector noise scales with the signal amplitude; consequently, both the signal and the noise increase proportionally with L, leaving the SNR approximately unchanged. As a result, the peak-to-concentration regression is mathematically invariant to path length, and the small variations originally observed in Figure 8 are attributable primarily to statistical sampling variability rather than to a genuine physical dependence [30].

Table 8. Illustrative sensitivity analysis of retrieval mean absolute error (MAE) vs. optical path length, under proportional-only and proportional-plus-fixed detector noise models

Path Length (cm)	MAE, Proportional-Only (ppm)	MAE, Fixed + Proportional (ppm)
0.05	0.0799 ± 0.0051	0.0885 ± 0.0051
0.10 (manuscript default)	0.0799 ± 0.0051	0.0825 ± 0.0051
0.20	0.0799 ± 0.0051	0.0806 ± 0.0051
0.50	0.0799 ± 0.0051	0.0800 ± 0.0051
1.00	0.0799 ± 0.0051	0.0799 ± 0.0051

In practical optical systems, however, detector electronic noise, dark-current noise, and digitization noise introduce approximately fixed contributions that do not scale with absorbance. To illustrate this physically expected behavior, an additional sensitivity analysis was performed by combining the original proportional noise model with a representative fixed detector-noise component. The results (Table 8) demonstrate the expected physical trend: retrieval accuracy deteriorates at short optical path lengths because the absorbance approaches the fixed noise floor, reducing the effective SNR.

As the optical path length increases, the signal grows while the fixed noise remains approximately constant, resulting in progressive improvement that gradually approaches a performance plateau. Under this illustrative analysis, the selected optical path length (L = 0.1 cm) already lies close to this plateau, remaining within approximately 3.3% of the performance obtained at L = 1.0 cm. This provides a physical justification for selecting a short optical path, since longer gas cells would provide only marginal improvement while increasing instrument size and complexity [28].

3.10 Receiver operating characteristic curve for diabetes screening

Test samples were then subdivided such that those >2 ppm were classified into a diabetic group, while those ≤2 ppm were classified into a non-diabetic group.

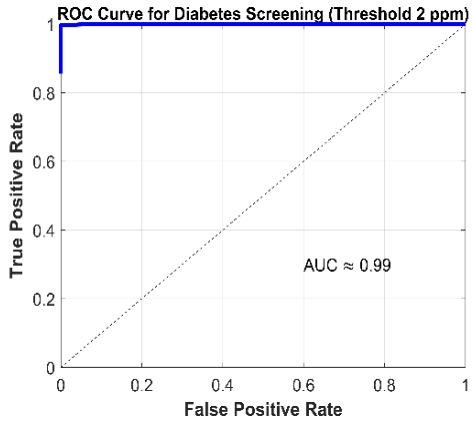


Figure 9. ROC curve for diabetes screening at the assumed 2 ppm threshold (AUC ≈ 0.99)
Note: Receiver operating characteristic (ROC), area under the curve (AUC).

The resulting ROC curve is plotted in Figure 9 and is convex toward the top-left corner. Trapezoidal integration of the area under the curve yielded an AUC ≈ 0.99, with minimal overlap between the two classes. The result reflects the discriminative separability of the simulated concentration distributions under the assumed 2 ppm threshold, rather than validated clinical diagnostic performance. Assessing true diagnostic performance would require evaluation against experimentally measured breath-acetone distributions from healthy and diabetic populations, which is identified as a priority for future work (Section 4.3).

3.11 Evaluation under a more realistic noise model

In order to assess the robustness of the proposed retrieval framework under realistic measurement conditions, the synthetic spectra were again generated from a more extensive noise model, as well as the original additive detector noise. As is common in infrared spectroscopy simulations, the following additional perturbations were introduced: multiplicative intensity fluctuations, linear baseline drift, and shifts along the wavenumber axis, which account for the uncertainty in

wavelength calibration. These perturbations are not hardware-specific but were chosen to represent some representative instrumental effects found in the spectroscopy literature.

To limit random sampling variability effects, the original and the enhanced noise models were analyzed 20 times over the same concentration, temperature, pressure and SNR realizations.

The complete dataset (2,000 spectra) was regenerated and the retrieval model was re-trained in each repetition.

Table 9. Performance comparison between the original and comprehensive noise models (20 paired repetitions)

Condition	R^2	MAE (ppm)	RMSE (ppm)
Original noise model	0.9821 ± 0.0023	0.5414 ± 0.0287	0.7684 ± 0.0507
Realistic noise model*	0.9801 ± 0.0021	0.5753 ± 0.0231	0.8092 ± 0.0477
Paired difference (realistic-original), mean ± standard deviation across 20 paired repetitions	−0.0019 ±0.0017	+0.0339 ± 0.0196	+0.0408 (of means)

Note: Coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE).
Realistic noise model*: the original additive detector noise was combined with multiplicative intensity noise (2% standard deviation), linear baseline drift (random slope, spanning up to 3% of the peak absorbance across the spectral window), and a wavenumber-axis shift (0.02 cm^{-1} standard deviation).

Table 9 summarizes the averaged performance obtained over the repeated simulations. Incorporating multiplicative intensity fluctuations, baseline drift, and wavenumber-axis shifts resulted in only a modest reduction in prediction performance compared with the original noise model. Overall, the observed changes remained small, indicating that the proposed retrieval framework is relatively insensitive to these representative instrumental perturbations under the assumptions adopted in this study. Nevertheless, the present evaluation remains simulation-based.

3.12 Benchmark comparison with previous studies

Table 10 shows the comparison between the proposed method and existing methods examined in computational models for acetone breath analysis. To date, two directions for research have emerged: partial least squares (PLS) methods, which can handle the complete spectrum, with the drawback of requiring multivariate calibration; deep-learning models, including artificial neural networks (ANNs) and convolutional neural networks (CNNs), which allow classification with more accuracy, but at the expense of the need for larger labelled datasets and specific frameworks [10, 11].

The framework uses linear regression on a single peak-absorbance feature, achieving MAE = 0.57 ppm and R^2 = 0.979 while remaining interpretable and requiring no MATLAB toolbox [8, 9].

The interpretability ranking in Table 10 is qualitative and reflects the transparency of the relationship between the measured spectral feature and the estimated concentration, rather than a standardized quantitative metric.

Table 10. Benchmarking comparison with previous studies (qualitative)

Approach	Typical Method	Data Requirements	Toolbox Dependency	Interpretability
This work	Linear regression on peak absorbance	Hundreds of spectra	None (core MATLAB)	High
Chemometric	PLS/PCR on full spectrum	Hundreds of spectra	Statistics Toolbox	Moderate
Deep learning	ANN on raw spectra	Thousands of spectra	Deep Learning Toolbox + GPU	Low

Note: Partial least squares (PLS), principal component regression (PCR), artificial neural networks (ANNs), graphics processing unit (GPU).
Qualitative comparison of typical literature approaches; "Typical Method" and "Toolbox Dependency" reflect different aspects (methodology vs. implementation) and are not directly comparable across rows.

4. DISCUSSION

The Beer-Lambert law is itself linear in concentration, and this linearity underlies the retrieval's accuracy: peak absorbance is directly proportional to the product $\sigma_{max} \cdot C \cdot L$.

This linear relationship was consistent even when Gaussian noise was added to the spectra and pressure and temperature were varied across clinically realistic settings; R^2 did not fall below 0.98 under any of these conditions.

The quadratic alternative showed no measurable improvement over the linear fit; therefore, the concentration range tested in this study is adequately described by a linear relationship (Figure 7).

Unlike many previous studies on breath acetone sensing, which rely either on advanced optical instrumentation (e.g., cavity-enhanced spectroscopy or quantum cascade lasers) or on data-driven machine-learning models [8-11], the proposed framework adopts a simple physics-based linear retrieval strategy implemented using standard MATLAB functions. In addition to providing accurate concentration estimation, the framework includes prediction intervals, residual diagnostics, Durbin-Watson analysis, and ROC-based evaluation referenced to the clinically relevant 2 ppm threshold. Compared with representative ANN-based approaches, the proposed method achieves comparable prediction accuracy while requiring substantially fewer training samples and a considerably simpler computational workflow.

The theoretical basis of the approach is the linearity of the Beer-Lambert law itself in concentration, which permits a single physically interpretable feature—the peak absorbance—to carry the full retrieval signal without recourse to multivariate calibration or learned representations. This minimal design offers full parameter interpretability, since the fitted slope and intercept map directly onto known physical quantities, and complete statistical validation-prediction intervals, residual diagnostics, and Durbin-Watson statistic-linked by ROC analysis to the clinically motivated 2 ppm threshold used in the simulation.

The approach also requires substantially fewer training samples (1,400) than typical deep-learning alternatives, reflecting the reduced parameter count of a two-parameter linear model.

4.1 Robustness to environmental changes

The sensitivity sweeps shown in Figure 3 indicate that

temperature and pressure have little effect on the MAE. This finding is of practical relevance for real breath measurements, where both exhaled-air temperature and ambient pressure can vary. The dominant influence is noise. However, even at the lowest tested SNR of 20 dB, the MAE remained below 0.3 ppm, leaving ample room for relatively inexpensive photodetectors in a POC device.

In addition to the statistical evaluation, a formal uncertainty propagation analysis was performed to assess the influence of measurement uncertainties in temperature, pressure, and optical path length on the retrieved acetone concentration. The results show that, under the assumed operating conditions, the combined expanded uncertainty associated with these environmental parameters remains substantially smaller than the statistical prediction interval. This indicates that the overall retrieval uncertainty is dominated by detector noise and spectral variability rather than by uncertainties in the environmental parameters. Nevertheless, these uncertainty estimates are based on representative assumptions and should be refined once experimental hardware specifications become available [27].

4.2 Clinical implications

96% of the predicted concentrations fell within ±2 ppm of the true value (Table 7). The same 2 ppm threshold is the one more commonly employed in the literature to distinguish breath of diabetic from non-diabetic breath [5].

Combined with the ROC analysis (AUC ≈ 0.99), the results indicate strong discriminative power within the present simulation-based framework. As noted in Section 3.10, this reflects simulated discrimination under the assumed concentration threshold rather than validated clinical diagnostic performance.

Each inference completes in roughly 9 µs, which leaves ample headroom for a low-power microcontroller running inside a portable, battery-driven POC device [21].

4.3 Limitations and future work

It should be kept in mind that these findings have certain shortcomings. These spectra have not been measured but calculated from HITRAN cross-sections based on the Beer-Lambert law; thus, there are no real instrumental artefacts present. The original simulation did not account for spectral interference from other breath gases, notably water vapour and CO_2, likely present in an actual exhaled sample. This is a significant limitation, investigated quantitatively below.

The absorption cross-sections of H_2O and CO_2 were calculated using HITRAN line-by-line data in the same spectral window (2,615–3,300 cm^{-1}), applying a Lorentzian line-shape approximation at the reference condition (296 K, 1,013.25 mbar), where pressure broadening dominates over Doppler broadening. At representative exhaled-breath concentrations (H_2O ≈ 5.5%, CO_2 ≈ 4.5%), the peak absorbance contribution from H_2O exceeded that of acetone by a factor ranging from around 130 (at 20 ppm acetone) to over 26,000 (at 0.1 ppm acetone), while within the clinically-relevant range (≥2 ppm) the contribution from CO_2 remained less than that of acetone.

When both interferents were added to the simulated spectra, neither retrieval technique yielded satisfactory retrieval, even when the acetone feature was considered to be the global maximum or within a window at its nominal wavelength (2,970 cm^{-1}); this was because the H_2O absorption is not restricted to isolated lines, but is spread over a wide spectral window.

These findings indicate that a multi-feature regression approach restricted to the same spectral window is unlikely to resolve the interference effectively. Physical removal of water vapour upstream—for example, using a molecular-sieve or Nafion drying stage, as commonly employed in commercial breath-gas analyzers—is therefore a prerequisite for practical implementation; this mitigation strategy has not, however, been experimentally verified for the specific acetone-retrieval system considered here, and remains to be confirmed in future work.

Based on the present analysis, an approximate residual H_2O target below 5 ppm near the clinical decision threshold is recommended. CO_2 interference was found to be comparatively minor and is not expected to require similar mitigation [24, 25].

The usable spectral window is also bounded to the 3–3.8 µm range exploited here. Baseline drift, which any real laser-based device exhibits over a measurement session, was not corrected in an actual signal, since none was available; its effect was instead tested as a simulated perturbation. This enhanced simulation, incorporating multiplicative intensity fluctuations, baseline drift, and wavenumber-axis shifts, produced only a modest reduction in retrieval performance relative to the original additive-noise model, indicating reasonable tolerance to these representative instrumental effects.

However, the evaluation remains simulation-based, and experimental testing with calibrated gas standards and real exhaled breath measurements is still needed before practical performance can be fully established.

Experimental validation with exhaled breath samples, quantitative evaluation of cross-interference of co-existing exhaled gases, and population-based Monte Carlo analysis of breath acetone data sets, experimentally measured, will be addressed by future work, with the aim of assessing diagnostic classification uncertainty under realistic clinical conditions. Prior to clinical deployment, these steps must be taken [28].

5. CONCLUSIONS

A linear retrieval framework for exhaled acetone has been presented, based on a single feature—the peak absorbance of spectra simulated from HITRAN2020 cross-sections through the Beer-Lambert law. The fitted model achieved, on the held-out test set, $R^2 = 0.979$, MAE = 0.57 ppm, and RMSE = 0.82 ppm.

Beyond these headline figures, the present work has provided a layer of statistical characterization that is not commonly reported in the breath-analysis literature, including a 95% prediction interval, residual diagnostics, a three-factor sensitivity analysis, and a ROC curve at the assumed 2 ppm threshold (AUC ≈ 0.99), reflecting simulation-based discrimination rather than validated clinical performance. Inference requires approximately 9 µs per sample, supporting future integration.

The proposed framework retained high retrieval accuracy under an enhanced realistic noise model, which highlighted the robustness of the framework and placed an emphasis on further experimental evaluation based on real breath samples.

A GUM-based propagation of uncertainties analysis also showed that uncertainties for the environmental parameters are

relatively low compared to the total statistical prediction uncertainty under the assumed operating conditions.

The present work also includes a more general measurement-science characterization of the proposed framework, in addition to the retrieval accuracy. This comprises a GUM-compliant uncertainty budget, physical interpretation of the path-length function under the noise model chosen, insightful examples of water-vapor interference and quantification thereof used in the application of the sample-conditioning, and transparent presentation of simulation-derived metrological performance indicators such as LOD, LOQ, and dynamic range.

Together, these findings go beyond predictive capability: they highlight the main limitations of the current simulation model and identify the engineering parameters (drying, SNR, and optical path length) required for next-generation real-world applications. Experimental validation using actual exhaled breath samples—including an assessment of H_2O and CO_2 interference—remains essential before clinical deployment.

REFERENCES

[1] Miekisch, W., Schubert, J.K., Noeldge-Schomburg, G.F. (2004). Diagnostic potential of breath analysis—Focus on volatile organic compounds. Clinica Chimica Acta, 347(1-2): 25-39. https://doi.org/10.1016/j.cccn.2004.04.023

[2] Henderson, B., Khodabakhsh, A., Metsälä, M., et al. (2018). Laser spectroscopy for breath analysis: Towards clinical implementation. Applied Physics B, 124: 161. https://doi.org/10.1007/s00340-018-7030-x

[3] Drabińska, N., Flynn, C., Ratcliffe, N., et al. (2021). A literature survey of all volatiles from healthy human breath and bodily fluids: The human volatilome. Journal of Breath Research, 15(3): 034001. https://doi.org/10.1088/1752-7163/abf1d0

[4] Wang, C., Sahay, P. (2009). Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits. Sensors, 9(10): 8230-8262. https://doi.org/10.3390/s91008230

[5] Lee, L., Kim, Y., Chun, B.J., et al. (2023). Analysis of acetone absorption spectra using off-axis integrated cavity output spectroscopy for a real-time breath test. Current Optics and Photonics, 7(6): 761-765. https://doi.org/10.3807/COPP.2023.7.6.761

[6] Paleczek, A., Rydosz, A. (2022). Review of the algorithms used in exhaled breath analysis for the detection of diabetes. Journal of Breath Research, 16(2): 026003. https://doi.org/10.1088/1752-7163/ac4916

[7] Pangerl, J., Moser, E., Müller, M., et al. (2023). A sub-ppbv-level acetone and ethanol quantum cascade laser based photoacoustic sensor–characterization and multi-component spectra recording in synthetic breath. Photoacoustics, 30: 100473. https://doi.org/10.1016/j.pacs.2023.100473

[8] Gordon, I.E., Rothman, L.S., Hargreaves, R.J., et al. (2022). The HITRAN2020 molecular spectroscopic database. Journal of Quantitative Spectroscopy and Radiative Transfer, 277: 107949. https://doi.org/10.1016/j.jqsrt.2021.107949

[9] Harrison, J.J., Allen, N.D., Bernath, P.F. (2011). Infrared absorption cross sections for acetone (propanone) in the 3 μm region. Journal of Quantitative Spectroscopy and Radiative Transfer, 112(1): 53-58. https://doi.org/10.1016/j.jqsrt.2010.08.011

[10] Bajo-Fernández, M., Souza-Silva, É.A., Barbas, C., Rey-Stolle, M.F., García, A. (2024). GC-MS-based metabolomics of volatile organic compounds in exhaled breath: Applications in health and disease. A review. Frontiers in Molecular Biosciences, 10: 1295955. https://doi.org/10.3389/fmolb.2023.1295955

[11] Kochanov, R., Gordon, I., Rothman, L., Wcisło, P., Hill, C., Wilzewski, J. (2016). HITRAN application programming interface (HAPI): A comprehensive approach to working with spectroscopic data. Journal of Quantitative Spectroscopy and Radiative Transfer, 177: 15-30. https://doi.org/10.1016/j.jqsrt.2016.03.005

[12] Gordon, I., Rothman, L., Hargreaves, R., et al. (2026). The HITRAN2024 molecular spectroscopic database. Journal of Quantitative Spectroscopy and Radiative Transfer, 353: 109807. https://doi.org/10.1016/j.jqsrt.2026.109807

[13] Kumail, A., Wei, J., Wang, C., et al. (2025). Acetone gas sensors for noninvasive diabetes diagnosis: A comprehensive review. The Chemical Record, 25(11): e2500105. https://doi.org/10.1002/tcr.202500105

[14] Guerrero, J., Gamba, M. (2025). A review of theory and practical considerations of tunable diode laser absorption spectroscopy diagnostics. arXiv preprint arXiv:2512.18201. https://doi.org/10.48550/arXiv.2512.18201

[15] Rossi, J., Vainio, M. (2024). Calibration-free infrared absorption spectroscopy using cantilever-enhanced photoacoustic detection of the optical power. Photoacoustics, 40: 100655. https://doi.org/10.1016/j.pacs.2024.100655

[16] Li, Z., Jiang, D., Zhang, M., Li, J. (2023). Diode laser absorption spectroscopy for real-time detection of breath oxygen. Infrared Physics & Technology, 133: 104815. https://doi.org/10.1016/j.infrared.2023.104815

[17] Mamouei, M., Budidha, K., Baishya, N., Qassem, M., Kyriacou, P.A. (2021). An empirical investigation of deviations from the Beer-Lambert law in optical estimation of lactate. Scientific Reports, 11: 13734. https://doi.org/10.1038/s41598-021-92850-4

[18] Zifarelli, A., Giglio, M., Menduni, G., et al. (2020). Partial least-squares regression as a tool to retrieve gas concentrations in mixtures detected using quartz-enhanced photoacoustic spectroscopy. Analytical Chemistry, 92(16): 11035-11043. https://doi.org/10.1021/acs.analchem.0c00075

[19] Saalberg, Y., Wolff, M. (2018). Multivariate analysis as a tool to identify concentrations from strongly overlapping gas spectra. Sensors, 18(5): 1562. https://doi.org/10.3390/s18051562

[20] Georgeous, J., Albasha, L., Husseini, G.A. (2025). A review of acetone vapor sensors for non-invasive diabetes diagnosis and management: Technologies, challenges, and future directions. IEEE Access, 13: 121261-121283. https://doi.org/10.1109/access.2025.3584863

[21] Kokkotis, C., Moustakidis, S., Swift, S.J., et al. (2025). Artificial intelligence and machine learning in the diagnosis and prognosis of diseases through breath analysis: A scoping review. Information, 16(11): 968. https://doi.org/10.3390/info16110968

[22] Ait Lahcen, A., Rajendran, J., Slaughter, G. (2026). Advances and challenges in digitally connected point-of-care biosensing. Biosensors and Bioelectronics: X, 28: 100728. https://doi.org/10.1016/j.biosx.2025.100728

[23] Welp, L.R., Keeling, R.F., Weiss, R.F., Paplawsky, W., Heckman, S. (2013). Design and performance of a Nafion dryer for continuous operation at CO2 and CH4 air monitoring sites. Atmospheric Measurement Techniques, 6(5): 1217-1226. https://doi.org/10.5194/amt-6-1217-2013

[24] Maiti, K.S., Hauer, J., Roy, S. (2025). Is digital drying an effective method for infrared spectroscopy of gaseous biofluid? Analytical Chemistry, 97(44): 24568-24575. https://doi.org/10.1021/acs.analchem.5c04584

[25] Lin, Y., Manalili, D., Khodabakhsh, A., Cristescu, S.M. (2024). Real-time measurement of CH4 in human breath using a compact CH4/CO2 sensor. Sensors, 24(4): 1077. https://doi.org/10.3390/s24041077

[26] Zhang, T., Kang, J., Meng, D., et al. (2018). Mathematical methods and algorithms for improving near-infrared tunable diode-laser absorption spectroscopy. Sensors, 18(12): 4295. https://doi.org/10.3390/s18124295

[27] JCGM 100:2008. (2008). Evaluation of measurement data—Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology. https://doi.org/10.59161/JCGM100-2008E

[28] JCGM 101:2008. (2008). Evaluation of measurement data—Supplement 1 to the "Guide to the expression of uncertainty in measurement"—Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology. https://doi.org/10.59161/JCGM101-2008

[29] Currie, L.A. (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). Pure and Applied Chemistry, 67(10): 1699-1723. https://doi.org/10.1351/pac199567101699

[30] Lin, S., Chang, J., Sun, J., Xu, P. (2022). Improvement of the detection sensitivity for tunable diode laser absorption spectroscopy: A review. Frontiers in Physics, 10: 853966. https://doi.org/10.3389/fphy.2022.853966