Optics EXPRESS

Gaussian process regression for direct laser absorption spectroscopy in complex combustion environments

WEITIAN WANG, ZHENHAI WANG, AND XING CHAO[*]

Center for Combustion Energy, Department of Energy and Power Engineering and Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Tsinghua University, Beijing 100084, China

[*] *chaox6@tsinghua.edu.cn*

Abstract: Tunable diode laser absorption spectroscopy (TDLAS) has been proved to be a powerful diagnostic tool in combustion research. However, current methods for post-processing a large number of blended spectral lines are often inadequate both in terms of processing speed and accuracy. The present study verifies the application of Gaussian process regression (GPR) on processing direct absorption spectroscopy data in combustion environments to infer gas properties directly from the absorbance spectra. Parallelly-composed generic single-output GPR models and multi-output GPR models based on linear model of coregionalization (LMC) are trained using simulated spectral data at set test matrix to determine multiple unknown thermodynamic properties simultaneously from the absorbance spectra. The results indicate that compared to typical data processing methods by line profile fitting, the GPR models are proved to be feasible for accurate inference of multiple gas properties over a wide spectral range with a manifold of blended lines. While further validation and optimization work can be done, parallelly composed single-output GPR model demonstrates sufficient accuracy and efficiency for the demand of temperature and concentration inference.

1. Introduction

The application of laser spectroscopic techniques in combustion diagnostics has greatly promoted combustion research [1,2]. As representative of laser-based sensing methods in practical environments, tunable diode laser absorption spectroscopy (TDLAS) is a sensitive, *in situ*, non-intrusive measurement technique that is capable of determining thermodynamic properties of gases, including concentration, temperature, pressure and velocity, etc. [3–5]. Direct absorption spectroscopy (DAS) with linearly scanned wavelength is the most commonly employed TDLAS scheme [4]. Typically, the absorbance data is post-processed by background reduction and line shape fitting to infer the targeted physical parameters, such as the absorbance area and line shape halfwidth [5]. While most such sensing systems scan over one or a few absorption transitions to be each profile-fitted individually, a broader spectral interrogation range covering more absorption lines provides opportunities for simultaneous inference of multiple thermodynamic parameters, presumably with better accuracy. With the booming development of advanced broadband laser sources and modern spectroscopic methods, absorbance spectra of larger bandwidth and higher resolution can be measured, based on which real-time, quantitative information of multiple species and parameters can be inferred with proper post-processing approach [6–9]. Nevertheless, the key challenge of processing a much larger number of lines over the spectral range of concern lies in the much more complicated multivariable optimization process, rendering conventional data processing algorithms costly or unstable [8,10]. Furthermore, when temperatures and pressures are high, enhanced pressure broadening and increasingly swamped spectra due to strengthened hot bands largely exacerbate spectral line overlapping, making it even harder to fit for both

#425662 https://doi.org/10.1364/OE.425662
Journal © 2021 Received 22 Mar 2021; revised 30 Apr 2021; accepted 1 May 2021; published 25 May 2021

TDLAS and the broadband schemes [11,12]. Several advanced spectral processing algorithms have been developed in face of the above-mentioned problems to improve the inference speed and accuracy. Blume et al. [8] proposed a broadband fitting approach to determine methane concentration, in which a line set to be fitted is treated as a fixed combination with regard to a main absorption line. The fitting residual over the entire spectral range is optimized by adjusting the parameters of the main line and the remaining lines accordingly. This allows the fitting of several hundred of lines with affordable computational cost, but inevitably at the expense of increased regression uncertainty. Mironenko et al. [13] proposed an algorithm to infer gas temperature based on the expansion of the spectra into a series of orthogonal polynomials, and the modified spectra with the first three components subtracted are fitted. Such approach increases the robustness of the fitting process by minimizing the number of free variables, but comparisons between simulated and experimental spectra are to be conducted in each iteration, thus limiting its application in real-time situations. Rein et al. [14] employed a differential evolution algorithm and optimized all variables simultaneously, including temperature, concentrations, and pressure. Similarly, the optimization process depends on iterative process that ensures convergence and accuracy at the price of efficiency.

While DAS finds its application in the diagnostics of various practical environments including the highly challenging scenarios associated with combustion, spectrally-resolved line shape fitting for quantitative detection is only possible for known species whose specific quantum transitions are targeted. Within the realm of chemometrics, which also utilizes absorption data for chemical analysis [15], the lack of knowledge of the definitive content requires inference based upon spectra of known components, the process of which known as spectroscopic calibration. Such conception of problem naturally leads to the implementation of machine learning and deep learning algorithms, which boast of recent development and success in efficient classification and data regression based on prior knowledge from training [16–18].

Among the various machine learning models, Gaussian process regression (GPR) is proved to perform better for spectroscopic calibration in chemometrics thanks to its flexibility of model and nonparametric nature [19]. The probabilistic covariance of the Gaussian process model also offers an explicit measure of the inter-relationship among training sets and therefore provides a built-in assessment of the uncertainties of the model prediction. This not only allows the algorithm to handle noisy data via automatic statistical tuning, but also makes GPR more physically expressive than other "black-box" algorithms such as neural networks.

In real applications, the number of outputs from a model is usually more than one, i.e., multiple properties need to be determined simultaneously, for example, temperature, pressure and concentrations of combustion gas mixtures. However, conventional GPR models usually allow only one single output, which means that the interdependency and correlation between the outputs have to be neglected. Extension of the output dimension for machine learning models including the GPR is nevertheless non-trivial and is still being investigated [20–23]. The challenge of such extension is to mathematically express the interrelations among the outputs and to optimize efficiently. Among the developed models, the linear model of coregionalization (LMC) proposed by Goulard et al. [24] is one of the most commonly used for its simplicity and high performance. With regard to chemometrics, Wang et al. [25] proposed a reduced covariance multi-response Kriging model by assuming the covariance function as the single-output covariance multiplied by the covariance between outputs, which can be regarded as a modification to the LMC structure.

In seeking for an efficient, quantitative algorithm for multi-output spectral data processing that are otherwise approached by traditional point-by-point least-square fitting, we here propose parallelly composed single-output and multi-output GPR models based on LMC for absorbance spectra processing in combustion gas monitoring. The training set is generated with simulated absorbance curves under varied known conditions using spectroscopic database, and the constructed and tuned model then directly yields thermodynamic properties of the gases such as temperature and concentrations from the observed absorbance spectra. Performance of the single-output and multi-output regression models are compared. Results show that the GPR algorithm provides an intelligent regression model such that thermodynamic properties can be inferred from the spectrally-resolved absorbance data directly and accurately, circumventing the costly and unstable multi-line fitting process. With the much improved computational efficiency by avoiding regression iterations and simulations over each measurement cycle, the proposed method demonstrates important potential to enable real-time spectral data reduction and simultaneous multi-variate diagnostics desired in complex combustion environments.

2. Methods and model description

2.1. Direct absorption spectroscopy

The fundamentals of laser absorption spectroscopy are well detailed in previous works [4,5], and only essential basics are briefly reproduced here to define symbols and units. The Beer-Lambert equation describes the relation between the transmitted intensity I and incident intensity I_0 when monochromatic light of frequency ν [cm^{-1}] passes through an absorption medium,

$$T_\nu = (\frac{I}{I_0})_\nu = \exp{(-k_\nu L)} = \exp{(-\alpha_\nu)} \tag{1}$$

where T_ν [unitless] is the transmittance at frequency ν, k_ν [cm^{-1}] is the absorption coefficient as characteristic of the medium, L [cm] is the absorption path length, and α_ν is the absorbance.

For a certain absorption line associated with an individual absorption transition of the gas molecule, the absorption coefficient k_ν can be further expressed with three terms,

$$k_\nu = S\phi_\nu P_i \tag{2}$$

where S [cm^{-2}atm^{-1}] is the line strength, ϕ_ν [cm] is the line shape function, and P_i [atm] is the partial pressure of the absorbing species i [26].

In a typical scanned-wavelength DAS scheme, the narrow-linewidth laser at center frequency ν is tuned continuously across a targeted transition line such that spectrally resolved absorption feature can be reduced from the transmitted laser signal recorded with a photo-detector after proper baseline reduction schemes [27]. For broadband techniques, the used laser sources and spectral resolving techniques may be different, but the property treated absorbance spectra, encompassing absorption transitions of numbers up to thousands, can be fed into the subsequent data processing procedure in a largely similar way. Desired properties of the absorbing gas medium can then be backed out from the Beer-Lambert law with known spectroscopic attributes of the gas.

2.2. Single-output Gaussian process regression (SOGPR)

The classic SOGPR model can be readily understood from the function-space point of view as a regression model to recover the underlying process from noisy observed data in supervised learning [28], where the process function $f(\mathbf{x})$ with input $\mathbf{x}$ follows a Gaussian process (GP) distribution of mean function $m(\mathbf{x})$, taken as zero without loss of generality (assuming $f(x) \neq 0$ within a narrow variable range of x), and covariance function $\mathrm{Cov}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$ (also known as the kernel function):

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{3}$$

$\mathbf{x}$ and $\mathbf{x}'$ are two inputs that are d-dimensional vectors in a Euclidean space, and the kernel determines the kind of structure captured by the GP model. As a first approach, we use the

Optics EXPRESS

radial basis function (RBF) as the kernel function in this work with the embedded assumption for smoothness of the model function, which can be expressed as

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2l}|\mathbf{x} - \mathbf{x}'|^2\right) \tag{4}$$

where the L$_2$-norm $|\mathbf{x} - \mathbf{x}'|^2$ calculates the Euclidean distance between point $\mathbf{x}$ and $\mathbf{x}'$ in the d-dimensional space, the pre-reference factor σ_f^2 characterizes the amplitude of output variance, and l represents the characteristic length scale, depicting the distance for the input vector to move until the output value changes significantly. These parameters are referred to as the hyper-parameters, specifying a particular distribution over function parameters.

In the context of the absorption spectra processing, the input vector $X \in R^{n \times d}$ is taken to be the sampled array of absorbance, with size d denoting the input dimension, or the number of evenly sampled wavenumbers of each absorbance curve, and n the number of processed absorption curves. Note that the covariance is calculated across different absorbance curves, and hence, quantifies the similarity of absorbance spectra under different thermodynamic conditions. As a result, if there are n training sets under different conditions to form matrix $X \in R^{n \times d}$ and n_* test sets to form $X_* \in R^{n_* \times d}$, the covariance matrix $K(X_1, X_2)$ with $X_{1,2} = X$ or X_*, can be constructed from the kernel function with its ij-th element $K_{ij} = k(\mathbf{x}_i^1, \mathbf{x}_j^2)$, where $\mathbf{x}_i^1$ and $\mathbf{x}_j^2$ are the i-th and j-th row-vector element of matrix X_1 and X_2, respectively. Assuming the observations y have independent and identically distributed white Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, i.e., $y = f(x) + \epsilon$, the observation vector $\mathbf{y}$ of training set X and the prediction vector $\mathbf{f}_*$ of test set X_* follows a joint Gaussian distribution,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \tag{5}$$

By the formula of Gaussian conditionals, the predictive distribution of the function value $\mathbf{f}_*$ at test points X_* follows a Gaussian distribution, which can be obtained based upon the GP prior tuned by training data observations through Bayesian inference:

$$\begin{aligned} &\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{Cov}(\mathbf{f}_*)), \text{ where} \\ &\bar{\mathbf{f}}_* \equiv \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1}\mathbf{y} \\ &\mathrm{Cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1} K(X, X_*) \end{aligned} \tag{6}$$

The mean of this modified Gaussian distribution ($\bar{\mathbf{f}}_*$) is then the predictions at test inputs yielded by the GPR model, and the variance (main diagonals of Cov($\mathbf{f}_*$)) provides a direct measure of the uncertainty.

2.3. Multi-output Gaussian process regression (MOGPR) and LMC

The extension of the model to MOGPR concerns about extending the number of outputs to P, i.e., for each observation, $\mathbf{f} = [f_1, \ldots, f_P]^T$. Following similar procedure as SOGPR, the MOGPR model with zero mean value for each output is described in [22],

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_M(\mathbf{x}, \mathbf{x}')) \tag{7}$$

where the multi-output covariance function $k_M(\mathbf{x}, \mathbf{x}')$ is now a $P \times P$ matrix,

$$k_M(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} k_{11}(\mathbf{x}, \mathbf{x}') & \cdots & k_{1P}(\mathbf{x}, \mathbf{x}') \\ \vdots & \ddots & \vdots \\ k_{P1}(\mathbf{x}, \mathbf{x}') & \cdots & k_{PP}(\mathbf{x}, \mathbf{x}') \end{bmatrix} \tag{8}$$

and $k_{pp'}(\mathbf{x}, \mathbf{x}')$ is the covariance function between output number p and p' $(1 \leq p, p' \leq P)$. Similar to the single-output condition, the prediction of test set $\mathbf{f}_*$ follows the joint Gaussian distribution,

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{Cov}(\mathbf{f}_*)) \tag{9}$$

where the mean and covariance are

$$\bar{\mathbf{f}}_* = K_{M*}^T [K_M(X, X) + \Sigma_M]^{-1} \mathbf{y} \tag{10}$$

$$\mathrm{Cov}(\mathbf{f}_*) = K_M(X_*, X_*) - K_{M*}^T [K_M(X, X) + \Sigma_M]^{-1} K_{M*} \tag{11}$$

$X \in R^{n \times d}$ is the training set matrix formed by n d-dimensional input vectors, while $X_* \in R^{n_* \times d}$ is the test set matrix containing n_* row-vectors. $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_P]^T \in R^{nP \times 1}$ is the observations with independent and identically distributed white Gaussian noise $\boldsymbol{\epsilon}_p \sim \mathcal{N}(0, \sigma_{s,p}^2)$, $\mathbf{y}_p(\mathbf{x}) = \mathbf{f}_p(\mathbf{x}) + \boldsymbol{\epsilon}_p$, $p = 1, \ldots, P$, and $\Sigma_M = \sigma_{s,p}^2 \otimes I_n \in R^{nP \times nP}$, where ⊗ denotes tensor product. The extended covariance matrices for the multi-output case are

$$K_{M*} = \begin{bmatrix} K_{11}(X, X_*) & \cdots & K_{1P}(X, X_*) \\ \vdots & \ddots & \vdots \\ K_{P1}(X, X_*) & \cdots & K_{PP}(X, X_*) \end{bmatrix} \in R^{nP \times n_* P}, \tag{12}$$

$$K_M(X, X) = \begin{bmatrix} K_{11}(X, X) & \cdots & K_{1P}(X, X) \\ \vdots & \ddots & \vdots \\ K_{P1}(X, X) & \cdots & K_{PP}(X, X) \end{bmatrix} \in R^{nP \times nP} \tag{13}$$

$$K_M(X_*, X_*) = \begin{bmatrix} K_{11}(X_*, X_*) & \cdots & K_{1P}(X_*, X_*) \\ \vdots & \ddots & \vdots \\ K_{P1}(X_*, X_*) & \cdots & K_{PP}(X_*, X_*) \end{bmatrix} \in R^{n_* P \times n_* P} \tag{14}$$

It is noted that each covariance function $k_{pp'}$ is associated with a set of hyper-parameters and constitutes an optimization problem to solve. With the output dimension now extended to P, the number of covariance functions to optimize has increased from 1 to P^2 and the number of observation noise distributions also increases from 1 to P, thus making the optimization task more computationally heavy and challenging.

The LMC model is a solution to this complexity by expressing the output $f_p(\mathbf{x})(p = 1, \ldots, P)$ of each observation as a linear combination of Q latent Gaussian processes [22],

$$f_p(\mathbf{x}) = \sum_{q=1}^{Q} a_{pq} u_q(\mathbf{x}) \tag{15}$$

where $u_q(\mathbf{x}) \sim \mathcal{GP}(0, k_q(\mathbf{x}, \mathbf{x}'))$. The LMC assumes uncorrelated latent Gaussian processes, i.e. $\mathrm{Cov}[u_q(\mathbf{x}), u_{q'}(\mathbf{x}')] = 0$ for $q \neq q'$. Thus, the covariance function can be simplified,

$$\begin{aligned} k_{pp'}(\mathbf{x}, \mathbf{x}') &= \sum_{q=1}^{Q} \sum_{q'=1}^{Q} a_{pq} a_{p'q'} \mathrm{Cov}[u_q(\mathbf{x}), u_{q'}(\mathbf{x}')] \\ &= \sum_{q=1}^{Q} a_{pq} a_{p'q} k_q(\mathbf{x}, \mathbf{x}') \end{aligned} \tag{16}$$

Equivalently in matrix form,

$$k_M(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{Q} A_q k_q(\mathbf{x}, \mathbf{x}') \tag{17}$$

where $A_q \in R^{P\times P}$ is a positive semi-definite matrix known as the coregionalization matrix.

If the number of latent Gaussian processes is one, the LMC model reduces to intrinsic coregionalization model (ICM), where

$$k_M(\mathbf{x}, \mathbf{x}') = Ak(\mathbf{x}, \mathbf{x}') \tag{18}$$

such that the covariance functions between different outputs are all linearly correlated. Furthermore, if the non-principal diagonal elements of A are all zero, i.e., $A = diag(a_{11}, \ldots, a_{PP})$, the outputs are independent with each other and the ICM reduces to a model including P individual SOGPR models in parallel, referred to as parallelly composed SOGPR hereon.

2.4. Further discussion of the model

The principle of inference by either the GPR model or with further coregionalization including LMC or ICM is the conditional distribution of Gaussian process following (12) and (13), which can be interpreted as a linear regression of the input training points, or a nonlinear interpolation of the “look-up-table" (training set). The prediction is the mean distribution of the trained Gaussian process, and the covariance between the prediction point and training points is a measure of their similarities, quantified by the RBF kernel. Through training, the hyper-parameters of the model, such as characteristic lengths for each latent Gaussian process $u_q(\mathbf{x})$ [1 parameter for each $u_q(\mathbf{x})$], as well as terms of the coregionalization matrices ($2Pq$ for coregionalization matrices of LMC, $2P$ for ICM, and P for SOGPR) and Gaussian noises ϵ_p (1 for each output), are optimized simultaneously. Input points close to each other tend to have similar outputs, in other words, the GPR inference can be interpreted as a non-iterative comparison between the test and training spectra, and interpolation is conducted through conditional distribution of the Gaussian process.

The advantage of such GPR inference scheme compared to traditional look-up-table methods lies in the fact that the nonparametric nature of GPR allows incorporation of complexity and richness of the data without specifying a fixed functional form, and that the probabilistic approach yields the most probable characterization of data with confidence of prediction through conditioning, without the cumbersome iterative regression procedure for each new test point. Furthermore, the superiority of GPR over conventional least-squared-based optimization methods such as Levenberg-Marquardt algorithm lies in the iteration-free linear model structure. The most computationally expensive processes have been completed in the training step ahead of inferring, so each time given a new test set, only $K_{M*} \in R^{nP\times n_*P}$ is to be updated, yielding an $O(nn_*P^2)$ complexity. In contrast, the conventional optimization methods are iterative with a maximum $O(\epsilon^{-2})$ number of iterations, where ϵ is the tolerance much less than one [29]. Although it is not evident to compare the computational efficiencies of an iterative method and a linear method with analytical expression, it will be clearly shown from the following results that the GPR model performs better than conventional line shape fitting method, in terms of efficiency, accuracy, and stability. Such inference performance of GPR provides opportunities for fast, real-time spectroscopic diagnostics without human interventions and *ad hoc* model tuning.

3. Problem formation and model implementation

3.1. Dataset formation with molecular absorption spectroscopy

In this work, a gas mixture of H_2O, CO_2 balanced with air is investigated, and the temperature and concentrations of the species is considered as model outputs. The temperature is assumed to be uniform and pressure is assumed to be 1atm. Therefore, there are 3 outputs in total: concentrations of H_2O and CO_2, and temperature of the gas mixture.

The simulation of absorbance spectra is conducted with HITEMP 2010 database [30] and the HITRAN Application Programming Interface (HAPI) [31]. The targeted wavenumber range is 3770 ~ 3780cm^{-1} in the mid-infrared, comprising 208 lines of CO_2 and 111 lines of H_2O. As shown in Fig. 1, these lines are densely spaced, and have comparable line strengths at high temperatures. The absorption length is set as L = 5cm.

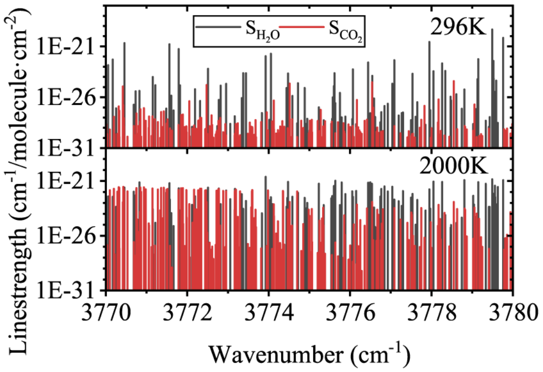


Fig. 1. Line strengths of 319 lines in 3770 ~ 3780cm^{-1} for H_2O and CO_2 under different temperatures.

The training set is formed in the 3-D vector space spanned by mesh grids of 1000 total points with 10 equidistant points for each parameter within ranges of $T = [1550, 2500]$K, $C_{H_2O} = [0, 0.09]$ and $C_{CO_2} = [0, 0.09]$, respectively, and the test set consists of 50 spectra at randomly drawn conditions from the same space. All generated absorbance spectra in the test set are distorted with independent and identically distributed white Gaussian noise with 1×10^{-5} absolute standard deviation. Figure 2 shows three of the simulated spectra in test set and their associated conditions.

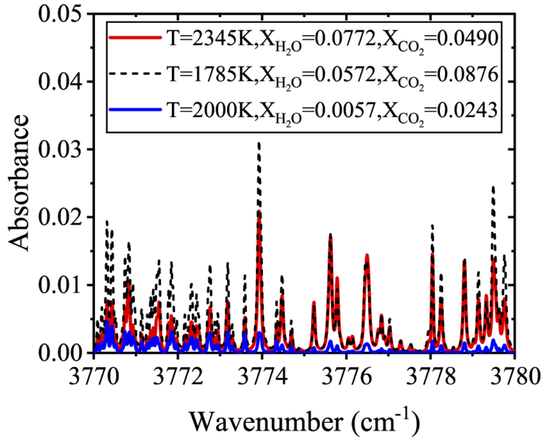


Fig. 2. Three absorbance spectra from the test set with randomly generated conditions.

3.2. Establishment of GPR models

The simulated training and test sets under different concentration and temperature conditions are used as input vectors of the GPR model, which is programmed to infer the two concentrations and temperature directly from the absorbance spectra. In accordance with such model formulation, the input vector is then the absorbance vector $(\alpha_{\nu_1}, \ldots, \alpha_{\nu_d})^T$ composed of the simulated absorbance at each wavenumber grid point. The output vector is the three thermodynamic properties $(T, C_{\mathrm{H_2O}}, C_{\mathrm{CO_2}})^T$. The models are implemented using GPy [32], which is a Gaussian process framework in python supporting both SOGPR and MOGPR based on LMC (with $Q = 3$) and ICM. RBF kernels are utilized in all models. Simulated training spectra are fed into the model to optimize the hyper-parameters, and test sets are validated using the trained model. Figure 3 illustrates the flowchart of the training and validation process.

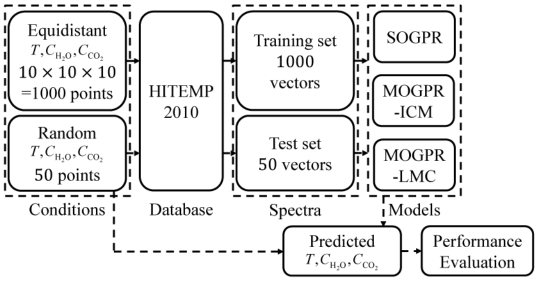


Fig. 3. Flowchart of the GPR validation process. SOGPR: single-output Gaussian process regression; MOGPR: multi-output Gaussian process regression; ICM: intrinsic coregionalization model; LMC: linear model of coregionalization.

Several preparational procedures are necessary before training. Since the absolute values of temperatures and concentrations differ significantly in their orders of magnitude, the physical parameters are normalized before being fed into the model for enhanced numerical robustness. Likewise, the pre-reference factors for all three RBF kernels are set to one, because the amplitudes of output variances are scaled by diagonal terms of the coregionalization matrix. The Gaussian noise term Σ_M may also be constrained to physically reasonable values to avoid getting trapped in local extrema in the optimization problem [10].

3.3. Training and validation results

The training process of the GPR boils down to optimizing the kernel parameters with constrains of the training set. A parallelly composed SOGPR model (by forcing coregionalization matrix A to be diagonal) as well as two MOGPR models including an LMC model and an ICM are trained using identical training set, and their performances on the test set are reported and discussed in the following. Figure 4 shows one of the spectra in Fig. 2 and the predicted spectrum reproduced using inferred parameters by ICM, together with the result from conventional line shape fitting. The residuals from both procedures are also reported, with the residual of GPR inference amplified by 100 times for clarity of comparison. It can be obviously concluded that the ICM provides more accurate and stable inference of the unknown spectrum.

Table 1 lists the root-mean-square error (RMSE) for simultaneous predictions of temperature, H_2O and CO_2 concentration by the three models, and Fig. 5 illustrates the predictions and errors for each point in the test set. It can be seen that all three models provide accurate predictions for each thermodynamic property with relative RMSE of about 1% for temperature, about 0.5% for H_2O concentration, and about 2.5% for CO_2 concentration. The reason for a relatively larger error for CO_2 concentration prediction attributes to the relatively lower line strengths of CO_2 at lower temperature, which can be seen from Fig. 1. In Fig. 5, we notice that there is one outlier

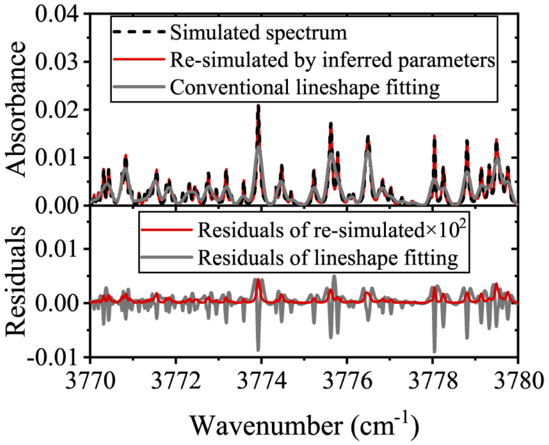


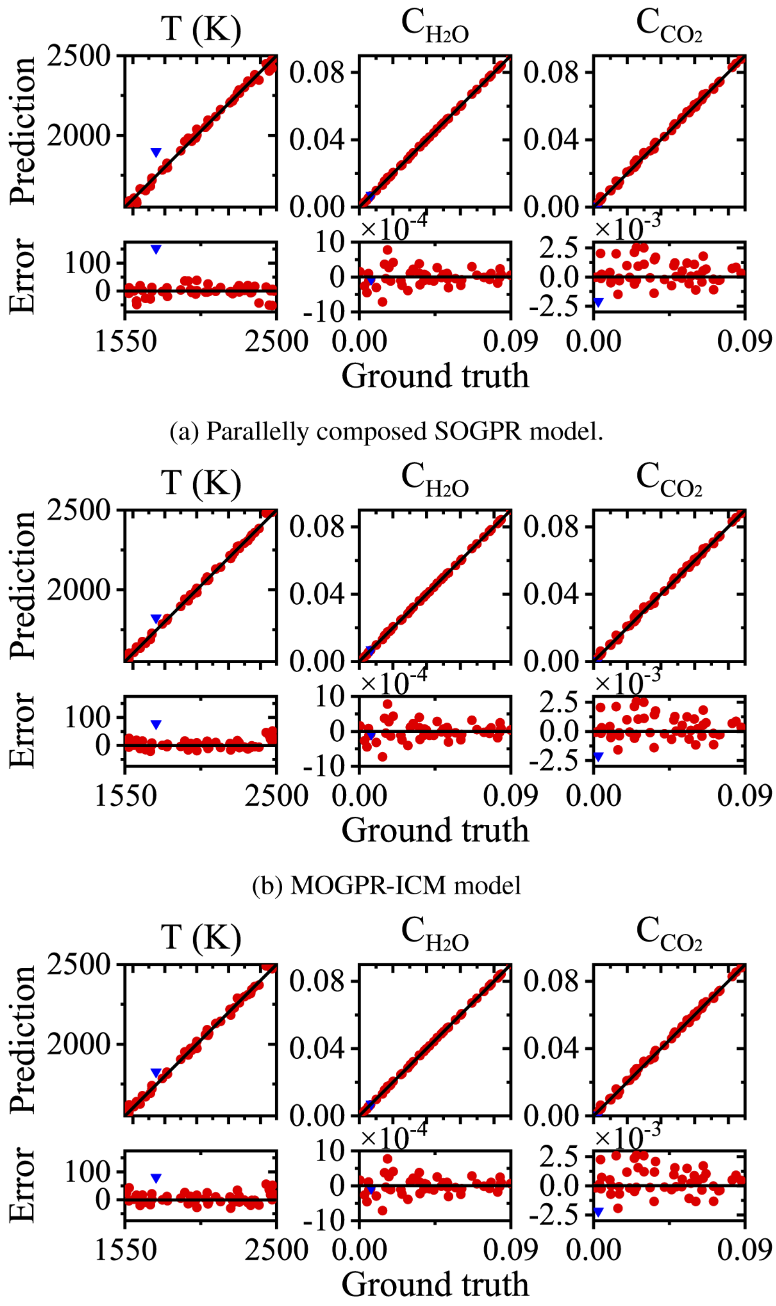
Fig. 4. Spectrum inference of ICM by re-simulating using inferred parameters, and multi-line shape fitting as well as their residuals. The residuals of the former are amplified by 100 times.

point (marked by blue triangle) with a large relative prediction error of 8.77% for temperature that stands out on all three temperature prediction figures. This is associated with a low H_2O and CO_2 concentration condition, with T = 1745.68K, C_{H_2O} = 0.0072 and C_{CO_2} = 0.0028. Under such condition, the absorption is weak (maximum absorbance $\alpha_{max} = 4.44 \times 10^{-3}$) and distorted by the white noise, which degrades the prediction accuracy at this point. Figure 6 shows the relation of prediction errors against integrated absorbance (sum of absorbance over the wavenumber range of interest) for three models. As shown in Fig. 6, the absolute prediction errors for all outputs are relatively large in the region of integrated absorbance lower than 0.04cm^{-1}, especially when concentration is low and integrated absorbance is close to zero, which agrees with the case of the outlier point in Fig. 5. It is also worth noticing that there are abnormal increases of all three outputs when the integrated absorbance falls near its maximum, especially for CO_2 concentration. The explanation is that under such condition, all parameters approach their boundaries and the lack of information outside the given parameter range leads to an increase of uncertainty.

Table 1. Root-mean-square error (RMSE) and relative RMSE (relative to the mid-point values of 2025K, 0.045, 0.045, respectively) for predictions by each parallel SOGPR, MOGPR-ICM, and MOGPR-LMC model.

RMSE/relative RMSE	Parallel SOGPR	MOGPR-ICM	MOGPR-LMC
T(K)	29.50/1.46%	19.35/0.96%	22.51/1.11%
C_{H_2O}	2.41×10^{-4}/0.53%	2.47×10^{-4}/0.55%	2.41×10^{-4}/0.54%
C_{CO_2}	1.07×10^{-3}/2.38%	1.09×10^{-3}/2.43%	1.20×10^{-3}/2.67%

To understand the influence of parameters and identify the source of errors, an additional run of simulation with temperature fixed at 2025K and H_2O and CO_2 concentration grids same as above-mentioned is conducted. In essence, this reduces the number of training set variables to two, but keeps the output number at three. The constrained parameters yield better inference accuracy by ICM, with RMSE = 10.07K for temperature, and RMSE = 7.90×10^{-6} and 2.52×10^{-5} for H_2O and CO_2 concentrations, respectively. The surface in Fig. 7 shows the prediction errors of temperature on the training set containing 100 points. The conclusion mentioned above can be confirmed such that the prediction errors are relatively low in the middle of the parameter range, where the plane of ground truth (shown in gray) intersects the prediction surface. The errors then increase towards edges, especially near the origin where integrated absorbance is low. If



(a) Parallelly composed SOGPR model.

(b) MOGPR-ICM model

(c) MOGPR-LMC model

Fig. 5. Predictions and errors for each point by each model. Point marked by blue triangle is the outlier point discussed in Section 3.3

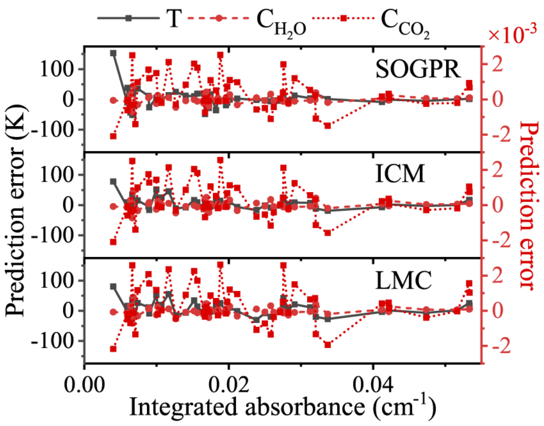


Fig. 6. Relation of prediction errors against integrated absorbance for each model.

we focus on a single contour of integrated absorbance (black lines on the surface), we notice that the error increases with the increase of CO_2 concentration. One plausible explanation is that the CO_2 transitions in the simulated range are mostly high-temperature lines, that are also temperature sensitive. As a result, the same disturbance leads to larger error under larger relative CO_2 concentration with identical integrated absorbance. In conclusion, the two major sources of errors include: weak absorption features disturbed by noises, and lack of information near the edges of simulation parameter range. In practical cases, the potential range of parameters needs to be considered, so that a wider simulating range can be applied to avoid extrapolation.

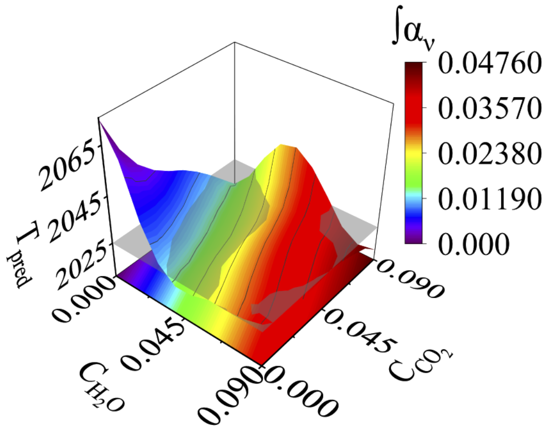


Fig. 7. Temperature predictions by ICM on constrained training set with fixed temperature at 2025K (marked by the gray plane). Colors with black contours correspond to integrated absorbance $\int \alpha_\nu$, instead of predictions.

3.4. Discussion of GPR model performance

Among the three models with increasing complexity, there is no clear evidence that the simplest parallel SOGPR model performs worse than the other two models, which is contrary to an intuitive speculation. The explanation is that a more complex model corresponds to more hyper-parameters (24 for LMC, 10 for ICM, and 7 for SOGPR in this situation), which complicates the optimization task and impairs model robustness. The validation shows that SOGPR model provides sufficient flexibility for absorbance inference, while LMC model and ICM can be more suitable for situations with more strongly correlated outputs. From another aspect, Fig. 8 (a) and (c) shows the common logarithm of covariance matrices $K_M(X, X) \in R^{30\times30}$ of the ICM and (b) and (d) shows that of the LMC model. Common logarithm of negative number is defined as opposite of the logarithm of its absolute value. As shown distinctly from the figures (a) and (b), both covariance matrices can be divided into nine $R^{10\times10}$ blocks; i.e., $K_M(X, X) = [K_M(X, X)_{i,j}],\ i, j = 1, 2, 3$. In this situation, the dominant terms of the covariance matrices locate on the main diagonal, while the values of the off-diagonal blocks are rather low, indicating weak correlations between the outputs. For the sub-block $K_M(X, X)_{1,1}$ in figure (c) and (d), it can also be seen that similar conditions correspond to larger covariance, which is consistent with expectation. As a result, we conclude that parallelly composed SOGPR model is sufficient for the studied case of temperature and concentration inference from the measured absorption spectra.

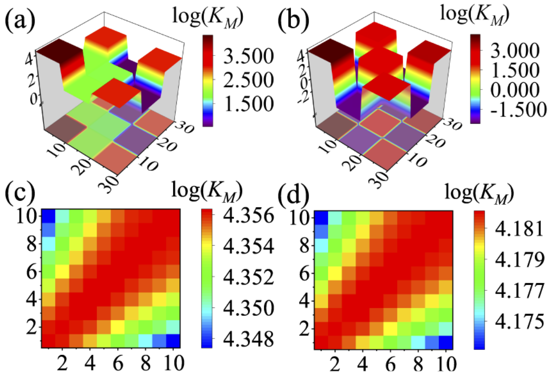


Fig. 8. Common logarithm of coregionalization matrices of the ICM (a) and (c) and the LMC model (b) and (d). (a) and (b) Three main diagonal blocks for each matrix correspond to the covariance of each output itself, and the other six off-diagonal blocks correspond to the covariance between outputs. (c) and (d) Zoom into the $K_M(X, X)_{1,1}$ block of each coregionalization matrix, which illustrates the covariance between prediction of one output (temperature for $K_M(X, X)_{1,1}$). Only 10 of training sets are utilized (T = 1550, 1656.6, . . . , 2500K, $C_{\mathrm{H_2O}} = C_{\mathrm{CO_2}} = 0.09$).

The Gaussian regression machine learning method for simultaneously inferring multiple thermodynamic properties from spectra shows irreplaceable advantage dealing with spectrally blended data that cannot be easily handled by traditional curve fitting methods. Compared to iterative look-up-table methods, the linear structure of GPR ensures accuracy while saving computation time of inference by pre-training. Nevertheless, two major drawbacks of the GPR method are worth mentioning. First, for the three outputs considered in this case, merely 10 points for each parameter form a training set of 1000 condition points in total, presenting a clear challenge for the optimization task in training. The total data size as the resolution of mesh grid raised to the power of the output dimension strongly limits the sample finesse of the training set, and consequently limits the prediction accuracy. In other words, there is a trade-off between accuracy and computational cost to determine the size of training set. The second problem is the challenge for obtaining accurate absorbance. Since the RBF kernel quantifies the difference of two points by their Euclidean distance, a bias of the distance leads to failure of inference. For example, if there are independent and identically distributed white Gaussian noises σ and σ' with variance s^2 overlaid on vectors $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$ and $\mathbf{x}' = [x'_1, x'_2, \ldots, x'_d]^T$, respectively, the expectation of their distance will increase by $\mathbb{E}[|(\mathbf{x}+\sigma) - (\mathbf{x}'+\sigma')|^2] - \mathbb{E}[|\mathbf{x}-\mathbf{x}'|^2] = 2ds^2$, which will bring increased error to the inference as the noise level and input dimension increase. Since the accuracy of absorbance strongly affect the accuracy of GPR models, proper baseline-reduction algorithms and denoising methods should be applied to real experimental data before the inference procedure. Attempts to incorporate these factors into more expressive kernels are also underway to make the GPR method more versatile.

3.5. Code availability

The code implementing the above-mentioned GPR models together with the simulated data can be accessed in Code 1 (Ref. [33]). GPy is required when running the script. Please follow the instructions of [32] to install GPy properly.

4. Conclusion

In this work, we develop and validate a machine learning model based on GPR to directly infer gas thermodynamic properties from direct absorption spectroscopy data in combustion environments. Once trained, both parallelly composed SOGPR model and MOGPR model are able to infer temperature and concentration of H_2O and CO_2 directly from the raw absorbance spectra accurately and efficiently without iterations, simulations in-the-loop, or human interventions, which makes real-time, unsupervised combustion diagnostics with high accuracy possible. The results demonstrate that for the case of inferring temperature and two concentrations, parallelly composed SOGPR model and MOGPR models including LMC model and ICM performs equally well, with relative RSME <3% for all three outputs, and the SOGPR model is recommended for its simplicity.

While this work develops and validates an effective data inference algorithm for multi-transition absorbance spectra, future work can be done in three aspects: First, the algorithm has been demonstrated through a challenging showcase wavelength region which is representative of two competitively strong absorbing species and more than three hundred lines. Immediate experimental applications for both high-temperature combustion gas sensing as well as broadband laser absorption spectroscopy are underway to exploit the non-iterative feature of the developed method. Second, the algorithm can be further developed to incorporate the pre-processing step that translates the raw measured signal into the absorbance spectra. This step will be tuned specifically to comply with the measurement technique and the associated baseline reduction schemes, such as neural network regression [34], cepstral analysis [9], and Bayesian inference [35,36]. Finally, improved kernel that is more expressive of the laser absorption signal can be constructed and developed to further increase the accuracy and robustness of the model.

Funding. National Natural Science Foundation of China (51976105, 61627804, 91841302); Seed Fund of Shanxi Research Institute for Clean Energy (20182000335); National Defense Basic Scientific Research Program of China (JCKY2019204B020).

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are available in Ref. [33].

References

1. O. Witzel, A. Klein, C. Meffert, S. Wagner, S. Kaiser, C. Schulz, and V. Ebert, "Vcsel-based, high-speed, in situ tdlas for in-cylinder water vapor measurements in ic engines," Opt. Express **21**(17), 19951–19965 (2013).
2. D. Wen and Y. Wang, "Spatially and temporally resolved temperature measurements in counterflow flames using a single interband cascade laser," Opt. Express **28**(25), 37879–37902 (2020).
3. J. Hodgkinson and R. P. Tatam, "Optical gas sensing: a review," Meas. Sci. Technol. **24**(1), 012004 (2013).
4. C. S. Goldenstein, R. M. Spearrin, J. B. Jeffries, and R. K. Hanson, "Infrared laser-absorption sensing for combustion gases," Prog. Energy Combust. Sci. **60**, 132–176 (2017).
5. Z. Wang, P. Fu, and X. Chao, "Laser absorption sensing systems: challenges, modeling, and design optimization," Appl. Sci. **9**(13), 2723 (2019).
6. M. J. Thorpe, K. D. Moll, R. J. Jones, B. Safdi, and J. Ye, "Broadband cavity ringdown spectroscopy for sensitive and rapid molecular detection," Science **311**(5767), 1595–1599 (2006).
7. L. Ma, W. Cai, A. W. Caswell, T. Kraetschmer, S. T. Sanders, S. Roy, and J. R. Gord, "Tomographic imaging of temperature and chemical species based on hyperspectral absorption spectroscopy," Opt. Express **17**(10), 8602–8613 (2009).

8. N. G. Blume, V. Ebert, A. Dreizler, and S. Wagner, “Broadband fitting approach for the application of supercontinuum broadband laser absorption spectroscopy to combustion environments,” Meas. Sci. Technol. **27**(1), 015501 (2016).
9. R. K. Cole, A. S. Makowiecki, N. Hoghooghi, and G. B. Rieker, “Baseline-free quantitative absorption spectroscopy based on cepstral analysis,” Opt. Express **27**(26), 37920–37939 (2019).
10. M. Lin, X. Li, W. Cai, S. Roy, J. R. Gord, and S. T. Sanders, “Selection of multiple optimal absorption transitions for nonuniform temperature sensing,” Appl. Spectrosc. **64**(11), 1274–1282 (2010).
11. J. M. Weisberger, J. P. Richter, R. A. Parker, and P. E. DesJardin, “Direct absorption spectroscopy baseline fitting for blended absorption features,” Appl. Opt. **57**(30), 9086–9095 (2018).
12. P. J. Schroeder, A. S. Makowiecki, M. A. Kelley, R. K. Cole, N. A. Malarich, R. J. Wright, J. M. Porter, and G. B. Rieker, “Temperature and concentration measurements in a high-pressure gasifier enabled by cepstral analysis of dual frequency comb spectroscopy,” Proc. Combust. Inst. **38**(1), 1561–1569 (2021).
13. V. R. Mironenko, Y. A. Kuritsyn, V. V. Liger, and M. A. Bolshov, “Data processing algorithm for diagnostics of combustion using diode laser absorption spectrometry,” Appl. Spectrosc. **72**(2), 199–208 (2018).
14. K. D. Rein, S. Roy, S. T. Sanders, A. W. Caswell, F. R. Schauer, and J. R. Gord, “Multispecies absorption spectroscopy of detonation events at 100 khz using a fiber-coupled, time-division-multiplexed quantum-cascade-laser system,” Appl. Opt. **55**(23), 6256–6262 (2016).
15. M. Baudelet, Laser Spectroscopy for Sensing: Fundamentals, Techniques and Applications (Elsevier, 2014).
16. T. Chen, J. Morris, and E. Martin, “Gaussian process regression for multivariate spectroscopic calibration,” Chemom. Intell. Lab. Syst. **87**(1), 59–71 (2007).
17. O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, and J.-P. Huvenne, “Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation,” Chemom. Intell. Lab. Syst. **96**(1), 27–33 (2009).
18. X. Zhang, T. Lin, J. Xu, X. Luo, and Y. Ying, “Deepspectra: An end-to-end deep learning approach for quantitative spectral analysis,” Anal. Chim. Acta **1058**, 48–57 (2019).
19. C. Cui and T. Fearn, “Comparison of partial least squares regression, least squares support vector machines, and gaussian process regression for a near infrared calibration,” J. Near Infrared Spectrosc. **25**(1), 5–14 (2017).
20. J. Park and S. Choi, “Multi-response gaussian process for multidisciplinary design optimization,” in *AIAA Aviation 2019 Forum* (2019, p. 2888.
21. M. A. Alvarez, L. Rosasco, and N. D. Lawrence, “Kernels for vector-valued functions: A review,” arXiv preprint arXiv:1106.6251 (2011).
22. H. Liu, J. Cai, and Y.-S. Ong, “Remarks on multi-output gaussian process regression,” Knowledge-Based Syst. **144**, 102–121 (2018).
23. H. Borchani, G. Varando, C. Bielza, and P. Larranaga, “A survey on multi-output regression,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **5**, 216–233 (2015).
24. M. Goulard and M. Voltz, “Linear coregionalization model: tools for estimation and choice of cross-variogram matrix,” Math. Geol. **24**(3), 269–286 (1992).
25. B. Wang, T. Chen, and A. Xu, “Gaussian process regression with functional covariates and multivariate response,” Chemom. Intell. Lab. Syst. **163**, 1–6 (2017).
26. R. K. Hanson, R. M. Spearrin, and C. S. Goldenstein, *Spectroscopy and optical diagnostics for gases* (Springer, 2016).
27. Z. Wang, P. Fu, and X. Chao, “Baseline reduction algorithm for direct absorption spectroscopy with interference features,” Meas. Sci. Technol. **31**(3), 035202 (2020).
28. C. K. I. Williams, *Gaussian processes for machine learning* (Taylor & Francis Group, 2006).
29. R. Zhao and J. Fan, “Global complexity bound of the levenberg-marquardt method,” Optim. Methods Softw. **31**(4), 805–814 (2016).
30. L. Rothman, I. Gordon, R. Barber, H. Dothe, R. Gamache, A. Goldman, V. Perevalov, S. Tashkun, and J. Tennyson, “Hitemp, the high-temperature molecular spectroscopic database,” J. Quant. Spectrosc. Radiat. Transfer **111**(15), 2139–2150 (2010).
31. R. V. Kochanov, I. Gordon, L. Rothman, P. Wcisło, C. Hill, and J. Wilzewski, “Hitran application programming interface (hapi): A comprehensive approach to working with spectroscopic data,” J. Quant. Spectrosc. Radiat. Transfer **177**, 15–30 (2016).
32. Sheffield Machine Learning, “GPy: A Gaussian process framework in python,” Github (2012) [accessed 10 May 2021] http://github.com/SheffieldML/GPy.
33. W. Wang, “Gpy sample code,” figshare (2021) [accessed 10 May 2021] https://doi.org/10.6084/m9.figshare.14248205.
34. S. Malek, F. Melgani, and Y. Bazi, “One-dimensional convolutional neural networks for spectroscopic signal regression,” J. Chemom. **32**(5), e2977 (2018).
35. J. Emmert, S. J. Grauer, S. Wagner, and K. J. Daun, “Efficient bayesian inference of absorbance spectra from transmitted intensity spectra,” Opt. Express **27**(19), 26893–26909 (2019).
36. J. Emmert, S. Wagner, and K. J. Daun, “Quantifying the spatial resolution of the maximum a posteriori estimate in linear, rank-deficient, bayesian hard field tomography,” Meas. Sci. Technol. **32**(2), 025403 (2021).