Technische Universität München
TUM School of Computation, Information and Technology

# The Generation and Application of Synthetic Photoacoustic Absorption Spectra for Machine Learning

Elisabeth Barbara Moser

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. Matthias Althoff

Prüfende der Dissertation:

1. Hon.-Prof. Dr. Frank Jenko
2. Prof. Dr. Rudolf Bierl
3. apl. Prof. Dr. Georg Groh

Die Dissertation wurde am 04.10.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology  am 29.03.2024 angenommen.

# The Generation and Application of Synthetic Photoacoustic Absorption Spectra for Machine Learning

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

**Doctoral Thesis**

| | |
|---|---|
| Author: | Elisabeth Moser (née Wittmann) |
| Supervisor: | Prof. Dr. Frank Jenko |
| Advisor: | Prof. Dr. Frank Jenko and Prof. Dr. Rudolf Bierl |
| Submission Date: | 22. September, 2023 |

Ich möchte mich ganz herzlich bei all den Menschen bedanken, die mich unterstützt und durch diese Zeit begleitet haben. Bei meiner Familie, die immer für mich da ist. Bei meinem Mann, der sowohl Euporie als auch Enttäuschungen mit aushalten musste. Bei meinen wunderbaren Arbeitskollegen / Freunden aus dem SappZ. Besonders bei Stefan, Flonnie, Flocko, Andi, Martin, Max, Tom, Lukas, Jonas und vor allem bei Simon, der mir immer mit Rat und Tat zur Seite stand um das Chaos in meinem Kopf zu sortieren. Auch bei meinen zwei Korrekturleser*innen, Matthias und Maike. Und natürlich bei der Hanns-Seidel Stiftung und dem BayWISS Verbundkolleg Digitalisierung, die mich finanziell untersützt haben und wunderbare Vernetzungsmöglichkeiten boten. Schließlich möchte ich mich auch bei Prof. Bierl und Prof. Jenko für die vertrauensvolle Zusammenarbeit und die gute Betreuung während meiner Disseration bedanken.

# Contents

# Abstract

Gas sensing is an essential field of research and industry, driving and enabling many processes and new research frontiers. A large field of gas sensing is absorption spectroscopy, as it allows fast and reliable quantification of multiple molecules simultaneously. Absorption spectroscopy reaches its limits when trace gas concentrations are under investigation, and intricate techniques have been developed to improve the sensitivity. One of those techniques is photoacoustic spectroscopy, which has an excellent potential for miniaturization but is typically applied toward only one gas component. Recent advances in laser technology allow the development of spectral photoacoustic measurements and thus enable multi-component quantification of trace gases, which is especially interesting in, for example, natural gas processing or medical breath analysis. To enable multi-component quantification in those challenging environments, machine learning methods have started to gain attention. However, the high cost of generating labeled data hinders the application and full usage of the advancements in machine learning.

This thesis addresses multiple application challenges of more data-intensive machine learning methods. First, the adaptation of measured absorption spectra to different environmental configurations is extended. Commonly used simulation frameworks only allow for the simulation of small molecules, while only measured data is available for larger molecules. This data has often not been collected at the required environmental configuration and therefore is of inferior quality. A deep learning framework is presented and evaluated for the adaption of measured absorption spectra to new pressure configurations. This allows using more adequate spectra during synthetic data generation and sensor development.

Secondly, a method to generate synthetic photoacoustic spectra from absorption spectra is presented. This method can estimate physically grounded parameters associated with the measurement setup from just a few measurements and then generate samples over a large range of concentrations. Two examples are presented using amplitude and wavelength-modulated photoacoustic data One is an application from medical breath analysis, and the other is for methane quantification in natural gas monitoring. The method also allows for the integration of the effect of non-radiative relaxation, which affects photoacoustic spectra.

Finally, the generated synthetic data is used in machine learning tasks for multi-component quantification. Here, it is compared to measured datasets or augmentations thereof over a range of machine learning techniques. This provides the first extensive comparison of synthetic spectra for machine learning models in gas absorption spectroscopy. While the advantages of synthetic spectra in machine learning only apply under certain conditions, the method holds great potential for complex multi-component spectra, which can only be interpreted by larger machine learning models.

# Kurzzusammenfassung in Deutscher Sprache

Die Gassensorik ist ein wichtiger Bereich in Forschung und Industrie, der viele Prozesse und Entwicklungen vorantreibt und ermöglicht. Ein großes Feld der Gassensorik ist die Absorptionsspektroskopie. Sie ermöglicht eine gleichzeitige, schnelle und zuverlässige Quantifizierung mehrerer Analyten. Aber die Sensitivität ist oft nicht ausreichend, um Spurengaskonzentrationen zuverlässig zu quantifizieren. Daher wurden verschiedene Techniken entwickelt um eine höhere Empfindlichkeit zu erreichen. Eine dieser Techniken ist die photoakustische Spektroskopie, welche ein großes Potenzial zur Miniaturisierung bietet, aber in der Regel nur für einzelne Gaskomponenten angewendet wird. Fortschritte in der Laserentwicklung ermöglichen nun spektrale photoakustische Messungen und damit die Quantifizierung von Mehrkomponenten-Spurengasen. Dies ist beispielsweise bei der Erdgasverarbeitung oder in der medizinischen Atemanalyse von besonderem Interesse. Um die Multikomponenten-Quantifizierung in diesen schwierigen Umgebungen zu ermöglichen, werden Methoden des maschinellen Lernens verwendet. Die hohen Kosten für die Aufnahme gelabelter Daten verhindert jedoch die Anwendung der Fortschritte im Bereich des maschinellen Lernens.

Diese Arbeit befasst sich mit mehreren Herausforderungen des Einsatzes datenintensiverer, maschineller Lernmethoden in der photoakustischen Gasspektroskopie. Zunächst wird die Methodik der Anpassung gemessener Absorptionsspektren an verschiedene Umgebungsparameter erweitert. Gängige Techniken erlauben nur die Simulation kleiner Moleküle, während für größere Moleküle meist Messdaten herangezogen werden müssen. Diese Daten sind oft nicht bei den erforderlichen Umgebungsparametern aufgenommen, was die Qualität der aus diesen Daten abgeleiteten Spektren beeinträchtigt. Es wird ein Deep Learning Framework zur Anpassung gemessener Absorptionsspektren an neue Druckkonfigurationen vorgestellt und evaluiert. Dies ermöglicht die Verwendung geeigneterer Spektren bei der Generierung synthetischer Daten und in der Sensorentwicklung.

Außerdem wird eine Methode zur Erzeugung synthetischer photoakustischer Spektren aus Absorptionsspektren vorgestellt. Mit dieser Methodik können physikalisch fundierte Parameter aus nur wenigen Messungen geschätzt und schließlich Spektren über einen großen Konzentrationsbereich erzeugt werden. Es werden zwei Beispiele mit amplituden- und wellenlängenmodulierten photoakustischen Daten vorgestellt. Die Methode ermöglicht auch die Integration des Effekts der nicht-radiativen Relaxation, die photoakustische Spektren beeinflusst.

Schließlich werden die generierten synthetischen Daten mit Methoden des maschinelles Lernens zur Quantifizierung von Multikomponentenproben aus zwei Anwendungsbereichen verwendet. Hier werden sie mit gemessenen Datensätzen oder deren Augmentierungen über eine Reihe von Methoden des maschinellen Lernens verglichen. Dies ist der erste umfassende Vergleich von synthetischen Spektren für Methoden des maschinellen Lernens in der Gasabsorptionsspektroskopie. Obwohl die Vorteile synthetischer Spektren beim maschinellen Lernen nur unter bestimmten Bedingungen zum Tragen kommen, birgt die Methode ein hohes Potenzial für komplexe Mehrkomponentenspektren, die nur durch anspruchsvollere Methoden des maschinellen Lernens interpretiert werden können.

*Part I*

# Introduction

# 1 Introduction

For centuries increased knowledge of the air surrounding us has ensured safety and fueled development. From canaries in coal mines to detect explosive gases to current-day fine-tuned process management in power to gas plants [1]. Or from the "fishy reek of advanced liver disease" to current precise breath analysis systems for early disease detection [2]: Gas sensing has enabled engineering advances and new research frontiers. During the Coronavirus Disease 2019 (COVID-19) pandemic, the use of $CO_2$ sensors for indoor air quality control peaked, and research in breath gas analysis techniques for early COVID-19 detection expanded [3]. Cheap, miniaturized sensors for single components like $CO_2$ and large, expensive, precise systems like gas chromatography coupled with mass spectroscopy employed in breath analysis are available. Between those two poles, accurate but cheaper multi-component sensing systems are needed to enable further developments, for example, breath analysis for a large group of patients during a pandemic [3], [4]. With the ability for precise, specific detection of multiple gases, optical gas sensing shows excellent potential to fill this gap [1], [5].

Optical gas sensing uses light absorption by gas molecules to measure the gas composition. The absorption pattern of each molecule is unique and thus allows differentiation and, with additional calibration, even quantification of the concentration of the analyte. This popular absorption gas sensing technique is used in various fields, ranging from environmental and process monitoring to medicine [1], [6], [7]. Each field and application holds its challenges and requirements, demanding specific designs in sensor development. With complex gas matrices, the absorption patterns of different molecules can overlap, leading to spectral interference. Simulations of the absorption spectra are typically employed during development to account for spectral overlaps. In addition to spectral interference, different sensing approaches and devices can lead to additional interference in exchange for, i.e., higher sensitivity or a broader spectral range. The development of a new sensor has to balance all those requirements.

To this end, Photoacoustic Spectroscopy (PAS), a subclass of absorption spectroscopy, has gained a lot of interest in sensor development. PAS is, in contrast to most absorption sensing technologies, a direct measurement technique and thus can achieve higher sensitivities. At the same time, it has excellent potential for miniaturization [8]. While most typical PAS sensors are targeted at one specific molecule, recent approaches have successfully combined laser tuning with machine learning approaches to detect multiple components simultaneously [9]. The output wavelength of a laser can thereby be tuned by altering the input current or temperature of the laser, thus one laser can cover a range of wavelengths. In PAS additional non-linear effects occur during the signal generation [10], making the quantification more demanding than from pure absorption spectra.

Often, machine learning is portrayed as an easy solution to those problems, allowing a fast and reliable suppression of the interferences [11]. Nevertheless, the amount of training data required for machine learning models is very high and currently hinders the integration of machine learning solutions in sensor development as no large datasets are available. Sensor-specific acquisition is cumbersome [11]. Recently, a few examples of using simulated spectral data to train absorption spectroscopy machine learning algorithms have emerged [12], [13]. While those works highlight the feasibility of this approach, no evaluation of the scopes and limits of synthetic data for machine learning in gas spectroscopy has been performed. In PAS, prior to this work, synthetic data has not been employed to train machine learning algorithms to the best of the author's knowledge. This is due to the additional effects during signal generation that are not incorporated in current simulations.

The main gaps in the current research covered by this thesis are summarized as follows:

- Insufficient simulation of more complex gas molecules absorption spectra at certain environmental configurations
- Lacking methods to transform absorption spectra to photoacoustic spectra in a physically plausible approach
- Evaluation of the use of synthetic data in gas spectroscopic machine learning scenarios

**Insufficient absorption simulation:** Synthetic gas absorption spectra are commonly obtained by line-data-based simulation approaches grounded in physical chemistry. Those are collected in line databases, for example, in the High-Resolution Transmission Molecular Absorption Database (HITRAN) [14] and provide simulation capabilities for more than fifty molecules at a broad range of environmental configurations. For the remaining, more complex molecules, measured absorption cross-sections are available. Those measured data are not easily transferred to other environmental configurations, such as a different pressure or temperature. While those measured cross-sections are sometimes used without adaption at a different configuration this can lead to falsified claims, especially in highly sensitive systems. A known method for adapting to different environmental circumstances are Estimated-Pseudo Line Lists (EPLLs) developed by Toon et al. [7], [15], which require much knowledge about and multiple measurements of the molecule in question. This thesis presents and evaluates a deep learning-based method for environmental adaption of measured cross-sections, which only requires one measurement and is independent of the molecule in question. Those reduced requirements allow for a broader application but also limit the method's applicability to only increasing pressure configurations as of now. Further improvements are outlined in this work to increase the capabilities of this approach.

**Photoacoustic simulation:** The main difference between typical absorption and photoacoustic spectroscopy lies in signal generation. While in absorption spectroscopy, the difference between the emitted light and the light collected after passing through a gas is considered, in photoacoustic spectroscopy, the amplitude and phase of a standing pressure wave are required. This pressure wave is generated by a modulated light beam, leading to molecular excitation by absorption and subsequent relaxation. This alternation in the system's energy causes the pressure wave which is measured. This process is further elaborated on in Section 3.4. This technique can detect far smaller concentrations because the signal can be amplified physically and computationally. In addition, the signal can be measured directly, not as in absorption spectroscopy as the difference of two signals. On the other side, the photoacoustic signal generation introduces additional artifacts from the modulation and non-linear effects, such as non-radiative relaxation. Those additional effects have previously not been considered for synthetic spectra, hindering the use of synthetic data for machine learning. Müller et al. [16] developed an algorithmic approach to compute the non-radiative relaxation efficiency. This approach is integrated with a simulation of the signal acquisition in this work to provide physically grounded synthetic photoacoustic spectra and parameters.

**Evaluation of synthetic data for machine learning:** While synthetic data is commonly used in computer vision or reinforcement learning to augment the training distribution of datasets [17, Chapter 3, 7], it has only recently been tapped into for gas absorption spectroscopy. Previous publications by Goldschmidt et al. [12] and Prischepa et al. [13] provide the first studies of the use of machine learning models - neural networks in those cases - trained on synthetic data for gas absorption spectroscopy. Nevertheless, they lack an evaluation of the impact of synthetic data on their models, as they did not compare them to models trained on measured data. An evaluation of the influence of synthetic compared to measured data is performed in this work for two photoacoustic sensing setups. This evaluation covers a range of machine learning methods.

This thesis aims to provide a starting point for integrating simulated absorption spectra with machine learning, focusing on photoacoustic gas sensor development. The current status and availability of absorption simulations and data are summarized. Techniques to adapt measured spectral data to environmental circumstances are explained, and a new approach for pressure adaption is presented. This new deep learning-based adaption technique can still be used when there is too little information for the other methods. All optical gas sensor developers require those simulation tools to account for spectral interference. In addition, simulated spectra can be adapted further to train machine learning algorithms. Here, a focus is placed on the adaption of synthetic spectra to PAS. The modeling of Amplitude Modulation (am) as well as Wavelength Modulation (wm) PAS signals is described, as well as an example for each measurement technique presented. This includes a methane sensor under challenging humidity conditions and a breath analysis PAS sensor tailored towards the simultaneous detection of acetone, ethanol, and water. The use of synthetic data combined with machine learning algorithms allows for a faster detection required for clinical implementation of this sensor. For both approaches, extensive comparisons of the use of synthetic versus measured training data depending on the underlying machine learning model are performed to outline the scopes and limits of this approach.

This thesis also explores and extends the use of simulation in gas sensor development. Using simulated, synthetic data as a basis for machine learning models is evaluated. Applying synthetic data in machine learning during (PAS) gas sensor development can enable new research areas as it allows for integrating machine learning approaches into the workflow. In addition, it speeds up the sensor development process, as no cumbersome generation of a large measured training dataset is required. This enables faster, better sensor designs when multi-component detection is needed, as in breath analysis, to provide affordable point-of-care devices during a pandemic.

## Structure of this Thesis

This thesis is partitioned into multiple sections and chapters. Following the introductory Part I, the state of the art and theory are described in Part II, followed by the experiments and results of this thesis in Part III. Finally, a summary and conclusion can be found in Part IV.

The state of the art described in Part II is partitioned into multiple chapters, each outlining one important topic of this thesis. First, the physical chemistry underlying the effect of molecular absorption is described in Chapter 2. Electronic, as well as vibrational and rotational energy transitions are described. The relation between line intensity and the physical process is explained. In addition, physical effects influencing the line shape are outlined, including natural, Doppler, collision broadening, and Dicke narrowing. Chapter 3 focuses on gas sensing and photoacoustic signal acquisition. First, the Lambert-Beer law, as the foundation of all absorption sensors, is described, and an overview of optical gas sensing techniques is presented. This is followed by an in-depth introduction to photoacoustic spectroscopy, including the physical process of signal generation. The different measurement setups like Quartz-Enhanced Photoacoustic Spectroscopy (QEPAS) and other techniques to increase the system's sensitivity are described. This also includes different modulation techniques. Finally, the non-spectral effects applicable to photoacoustic spectroscopy are described. This concludes the theoretical principles regarding physical chemistry. Next, the computational methods of absorption modeling are presented in Chapter 4. The most commonly used Voigt line shape model is presented, as well as the more extensive Hartman-Tran model, which includes the integration of more physical effects than the Voigt model. Next, an overview of the available line databases and the simulation code is given. The line-based modeling approach's limitations are outlined, and alternate measurement cross-section databases are presented. Finally, approaches to adapt spectra to other environmental configurations as well as ab initio spectral generation approaches are presented. An overview of machine learning in gas spectroscopy is presented in Chapter 5. This includes a broad introduction to machine learning and supervised learning algorithms for spectroscopists. The current use of machine learning in photoacoustic gas spectroscopy is reviewed, and an overview of standard techniques in the field of spectroscopy are presented, which differ from typical machine learning. Finally, the fields of variable selection and spectral machine learning are summarized.

Part III describes and evaluates the research and results from this thesis. In Chapter 6, the deep learning method to adapt cross-sections to different environmental conditions is described and evaluated. After a description of the methodology, including the dataset creation and the model architecture, the results on pressure adaption to fixed values for synthetic, in-dataset evaluation are presented. To evaluate the model on measurement data, an out-of-dataset evaluation on measured absorption cross-sections is presented and compared to the Estimated-Pseudo Line List (EPLL) technique. The model is extended to continuous pressure adaption, and its limits are outlined on the example of temperature adaption. The next chapter, Chapter 7, is focused on the use and generation of synthetic data in photoacoustic spectroscopy. After an outline of the simulation process for Photoacoustic Spectroscopy (PAS) synthetic data, the two case studies are presented. First, in Section 7.1, a Quantum Cascade Laser (QCL) based Amplitude Modulation (am) PAS sensor for human breath analysis for acetone, ethanol, water, and carbon dioxide is presented. The synthetic data generation process, including the estimation of measurement setup uncertainties and their integration into the modeling chain, is presented. This is followed by an extensive study of the influence of synthetic training data on eight different machine-learning algorithms. Finally, the application of synthetic data in combination with variable selection to reduce the measurement time is presented. The second case study in Section 7.2 describes a Wavelength Modulation (wm) QEPAS sensor for methane quantification under difficult humidity conditions. Here, the synthetic data generation process is extended by integrating the relaxation efficiency and wavelength modulation. This is followed by a similar evaluation of the effect of synthetic data as done for the previous case study.

In Part IV, the methods and results of this work are summarized in Chapter 8. Next, an outlook for future work in the covered areas is provided in Chapter 9. Finally, the conclusions divided into spectral adaption and the use of synthetic data are presented in Chapter 10.

*Part II*

# State of the art

# 2 Absorption Spectra

To successfully create synthetic spectra, and analyze and improve chemometric machine learning algorithms, a deep understanding of the physical processes behind the underlying data, the absorption spectra, is required. Therefore, the following chapter will provide an overview of the essential effects governing absorption spectroscopy.

Optical spectroscopy is a powerful technique to analyze gases, liquids, or solids to detect or quantify components and gain insights into molecular structures. It is based on the interaction of light with matter. Many possible interactions exist, including absorption, transmission, scattering, or reflection of the light. While this work will focus on absorption spectroscopy, thus the absorption of a light quantum (photon) by a molecule, other effects also yield great results in trace gas analysis and have a substantial potential for the application of spectral chemometrics.

The following sections will provide insights into the physical fundamentals of molecular absorption. The three possible categories of absorption that constitute any spectrum will be explained. The position, strength, and shape of absorption lines will be addressed. Since this work focuses on gas analysis, effects that only affect liquid or solid spectroscopy will not be discussed.

Note that some knowledge about the structure and nature of molecules and quantum theory is assumed. The interested reader might, therefore, refer to [18, chap. 11, 12].

## 2.1 Absorption Spectra

The process of absorption constitutes an energy transfer. The energy of a photon can be expressed as:

$$E_{\mathrm{ph}} = \mathrm{h} \cdot \nu_{\mathrm{ph}} \tag{1}$$

with h being Planck's quantum of action, also known as the Planck constant, and $\nu_{\mathrm{ph}}$ being the frequency of the photon. The photon frequency is in the context of absorption spectroscopy often given as the photon wavenumber $\tilde{\nu}_{\mathrm{ph}}$ for Infrared (IR) and the wavelength $\lambda_{\mathrm{ph}}$ for Ultraviolet (UV) and Visible (VIS) regions. Those can be converted by:

$$\nu_{\mathrm{ph}} = \tilde{\nu}_{\mathrm{ph}} \mathrm{c}_0 = \frac{\mathrm{c}_0}{\lambda_{\mathrm{ph}}} \tag{2}$$

with $\mathrm{c}_0$ being the speed of light in vacuum. The advantage of using the wavenumber is that it is directly proportional to the photon's energy. Thus, it displays as energy equidistant spacing in spectral plots, leading to an x-axis that is also directly proportional to the energy.

The wavelength regions of light are visualized in Figure 1 with their wavelength. The UV, VIS, and IR regions are indicated. The UV region can be further divided into the UVC region from 10 – 280 nm, the UVB region from 280 – 320 nm, and the UVA region from 320 – 400 nm [20, Page 2]. Due to its smaller size, the VIS region is normally not divided further for molecular and atomic spectroscopy. The IR region again is typically subdivided into three regions according to ISO 20473:2007 [21]. The

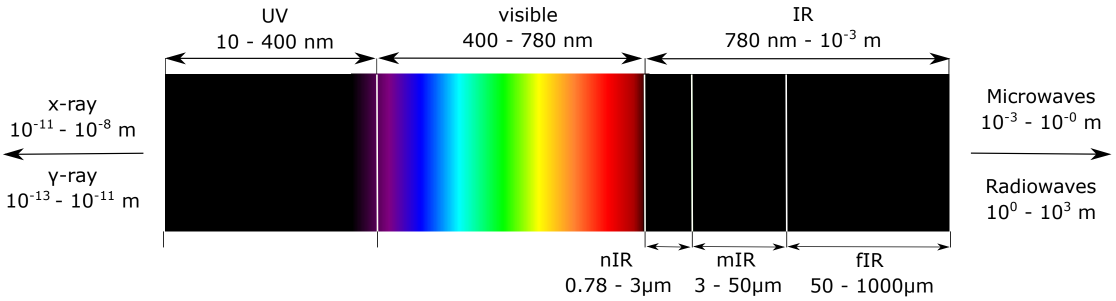


Figure 1: The wavelength regions of electromagnetic waves, important in molecular and atomic spectroscopy (figure adapted from [19, Page 43]).

Near-infrared (nIR) stretches from 0.78 – 3 $\mu$m (= 12 820.5 – 3 333.3 cm$^{-1}$, in wavenumbers). This is the most commonly used region for molecular gas analysis as most molecules possess absorption bands in this spectral region [22, Page 336]. It is also known as the fingerprint region. The region from 3 – 50 $\mu$m (= 3 333.3 – 200 cm$^{-1}$) is defined as the Mid-infrared (mIR) region while the Far-infrared (fIR) reaches from 50 – 1 000 $\mu$m (= 200 – 10 cm$^{-1}$).

Absorption of a photon can happen with a certain probability when its energy is equal to the energy difference of two energy states of a molecule. This results in a promotion of the molecule into a higher energy state. Three types of transitions are possible and will be explained in the next sections:

- electronic transitions
- vibrational transitions
- rotational transitions

Combinations of those basic transitions are also possible depending on the energy the photon provides. Interactions of all three basic transitions are called a molecules rovibronic transition, and a combination of the rotational and vibrational transition is called the molecules rovibrational transition.

### 2.1.1 Electronic Transitions

Electronic transitions will be denoted as $E_e$ and have the highest energy gap from the three transitions mentioned before, with $E_v$ for the energy of a vibrational and $E_r$ for the energy of a rotational transition.

$$E_e > E_v > E_r \tag{3}$$

The most famous electronic transitions are known as the Fraunhofer lines, first observed in 1814 by Joseph Fraunhofer in the spectrum of sunlight [23]. These dark regions in the spectrum correspond to transition lines of electronic transitions of several different molecules in the atmosphere. Under ambient conditions, molecules are usually in their electronic ground state. The energy needed to promote the molecules to a higher electronic energy level can only be provided by UV and VIS photons. Therefore, they are not visible in IR spectra [18, Page 557].

Not all electronic transitions are allowed, and selection rules apply. Since this work is focused on IR spectroscopy where electronic transitions don't occur, the inclined reader is referred to [18, chap. 17] for more information on electronic transitions, their selection rules, and relaxation pathways.

### 2.1.2 Rotational Spectra

Pure rotational spectra can only be excited in microwave spectroscopy, where they are used for chemical analysis like the determination of binding lengths [18, Page 531]. In IR spectroscopy, at least rovibrational, in VIS and UV spectroscopy, rovibronic states are excited. Those consist of an interaction of rotational and vibrational or rotational, vibrational, and electronic transitions, respectively.

To compute the energy of a certain transition of rotational states, the rotation constant $\tilde{B}$ is usually used [18, Page 524]:

$$\tilde{B} = \frac{\mathrm{h}}{8\pi^2 \mathrm{c}_0 I} \tag{4}$$

where I is the moment of inertia.

Most rotations can be computed with the help of classical physics and selection rules for allowed transitions from quantum mechanics. Only polar molecules exhibiting a permanent dipole moment show a pure rotation spectrum. Besides, the selection rule $\Delta j = \pm 1$ applies to all topologies. Here, $j$ denotes the total angular momentum quantum number. This means the rotation energy level can only be excited to the next higher level. In addition, the rotational energy levels can be distorted by different effects like centrifugal extension and degeneration. Those special distortion effects, a further discussion of selection rules, and examples for the computation of the rotational energy states for different topologies are given in [18, Page 522 ff].

### 2.1.3 Vibrational Spectra

In the IR spectral region, rotational and vibrational energy transitions are excited. Those will be explained separately first, even though, due to their smaller energy gap, rotational transitions are typically observed in combination with vibrational transitions. Those rovibrational transitions will be further explained in Chapter 2.1.4.

In its simplest approximation, a molecule can be regarded as a spring system where each connection of the atoms is approximated by a spring. The potential curve $V(x)$ corresponds to that of a harmonic oscillator, which has the shape of a parabola:

$$V(x) = \frac{1}{2} k_v x_r^2 \tag{5}$$

where $k_v$ is the spring force constant and $x_r$ is the atoms' bond length displacement. Considering the Schrödinger equation [24], this leads to vibrational energy levels given by:

$$E_v = (v + \frac{1}{2}) \frac{\mathrm{h}}{2\pi} \omega \quad \text{with} \quad \omega = (\frac{k_v}{m_{\mathrm{eff}}})^{\frac{1}{2}} \quad \text{and} \quad v = 0, 1, 2, ... \tag{6}$$

with $v$ being the vibrational quantum number and $m_{\text{eff}}$ the effective mass of the molecule. This effective mass for a diatomic molecule is given by $m_{\text{eff}} = \frac{1}{m_1} + \frac{1}{m_2}$ with the mass $\mathrm{m}_1$ and $\mathrm{m}_2$ of the two atoms constituting the molecule. The description of vibrational states can thus be simplified to:

$$G(v) = (v + \frac{1}{2})\tilde{\nu} \quad \text{with} \quad \tilde{\nu} = \frac{\omega}{2\pi \mathrm{c}_0} \tag{7}$$

This leads to a constant energy gap between each allowed state of $\Delta E_v = \frac{h}{2\pi}\omega$. However, this harmonic oscillator approximation neglects the anharmonicity, which significantly affects vibrational transitions, especially at higher vibrational quantum numbers.

This anharmonicity is rooted in the fact that molecules are actually not harmonic oscillators but can dissociate given a certain potential energy. This can be modeled by using the Morse-potential, which accounts for this anharmonicity [25]:

$$V(x) = \mathrm{hc}_0\tilde{D}_e[1 - e^{-a_M x_r}]^2 \quad \text{with} \quad a_M = \left(\frac{m_{\text{eff}}}{2\mathrm{hc}_0\tilde{D}_e}\right)^{\frac{1}{2}} \omega \tag{8}$$

with $a_M$ as a displacement factor and $\tilde{D}_e$ as the depth of the potential well corresponding to the energy needed for the dissociation. The anharmonicity of the potential curve is visualized against the curve of a harmonic oscillator in Figure 2. Combining the Morse potential with Equation 7, we arrive at the description of vibrational states as

$$G(v) = (v + \frac{1}{2})\tilde{\nu} - (v + \frac{1}{2})^2\tilde{\nu}x_e \quad \text{with} \quad v = 0, 1, 2 \tag{9}$$

with $x_e$ being the anharmonic constant

$$x_e = \frac{a_M^2 \mathrm{h}}{4\pi m_{\text{eff}}\omega} = \frac{\tilde{\nu}}{4\tilde{D}_e} \tag{10}$$

This leads to a finite number of possible states $v = 0, 1, 2, ..., v_{max}$, and the energy gap between different states gets smaller with higher vibrational quantum numbers. For vibrational transitions to be visible in the IR spectrum, the molecule's dipole moment needs to change with the transition. This entails that typically diatomic molecules of the same atom (like $N_2$ or $O_2$) are not IR active since their dipole moment does not change with vibrational transitions. In addition, the selection rule of $\Delta v = \pm 1$ applies to vibrational transitions. This selection rule is, however, not strict since it only applies to harmonic oscillators. Due to the anharmonicity of the vibrations, bigger transitions are partially allowed and are called overtones with $\Delta v = +2$ and higher. They only appear at a much smaller intensity, though.

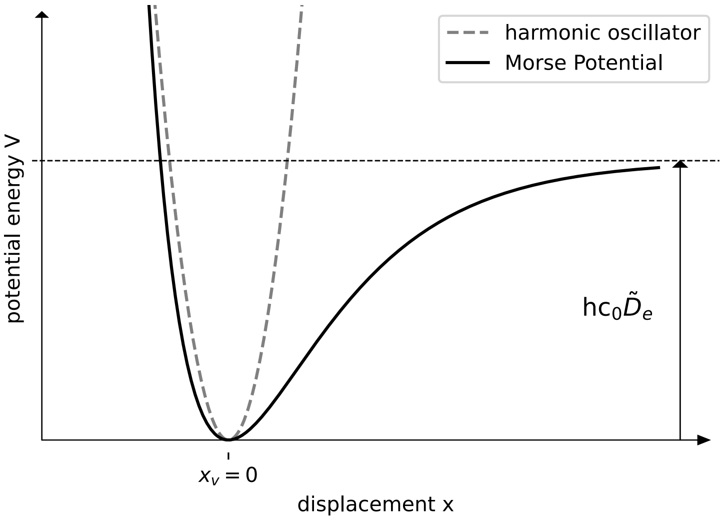


Figure 2: The potential of an harmonic oscillator (Equation 5) and the Morse potential (Equation 8) visualizing the potential energy V against the displacement x. This shows the anharmonicity of molecular vibrations at higher energies. The depth of the potential well $hc_0\tilde{D}_e$ is also visualized.

Bigger molecules of more than two atoms can be excited by different types of vibrations. The number of vibrational degrees of freedom can be computed for non-linear molecules by $3N_{Atoms} - 6$ and for linear molecules by $3N_{Atoms} - 5$. [18, Page 543]. Vibrations of molecules can be categorized into the basic categories of [20, Page 47]:

- stretching vibrations, which change the bond lengths within the molecule
- deformation vibrations, which alter mainly the bond angles.

Another classification scheme regards the symmetry of the vibrations. This is further described in [18, Page 544 ff].

### 2.1.4 Rovibrational Spectra

As already mentioned in Chapter 2.1.3, vibrational transitions are typically only observed in combination with rotational transitions [18, Page 539]. Such a spectrum is shown in Figure 3. Here a simulated absorption cross-section of $CO_2$ is presented. The spectrum was created with data from the High-Resolution Transmission Molecular Absorption Database (HITRAN) database [14] and the Voigt simulation algorithm [26] in HITRAN Application Programming Interface (hapi) [27]. This simulation process will be further described in Chapter 4. This spectrum shows a typical structure of a rovibrational transition, which displays a P, Q, and R–branch marked in the spectrum. This structure results from the selection rule of a rovibrational spectrum, which only allows transitions with $\Delta j = \pm 1$ for all molecules, resulting in the P–branch with $\Delta j = -1$ and the R–branch with $\Delta j = +1$. $CO_2$ also shows a typically forbidden transition of $\Delta j = 0$, which results in a strong Q–branch. Their exemplary transitions are visualized in Figure 4. The Q-branch transition is only allowed for a few certain transitions due to the molecule's symmetry. Therefore, many spectra typically show a space of the size $4\tilde{B}$ between the P and R–branches. The peaks within the branches have an equidistant spacing of $2\tilde{B}$. The different intensities visible for the different branches will be discussed in Chapter 2.2.

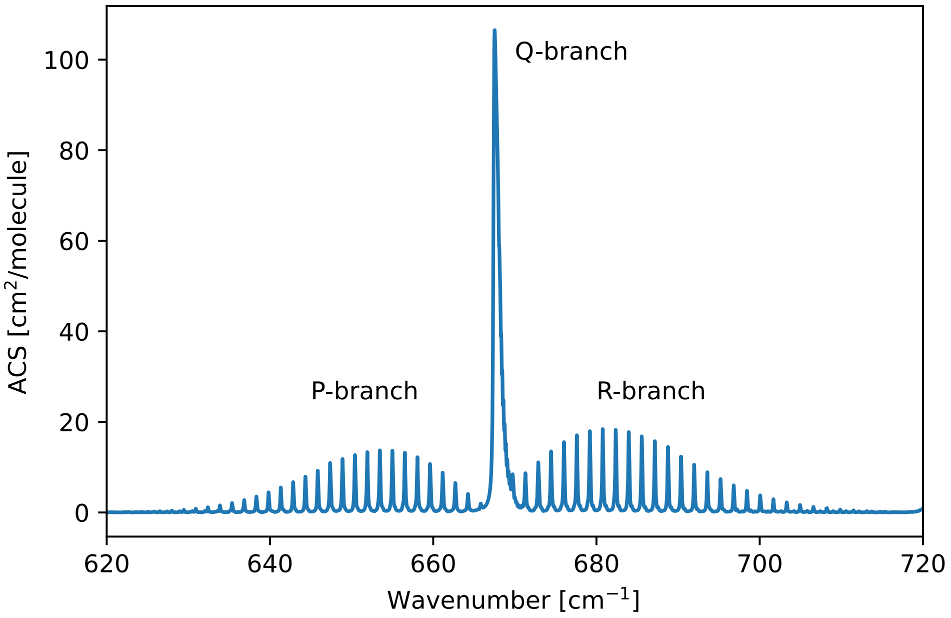


Figure 3: The simulated absorption cross-section of $CO_2$. The differences between P, Q, and R-branch are indicated.

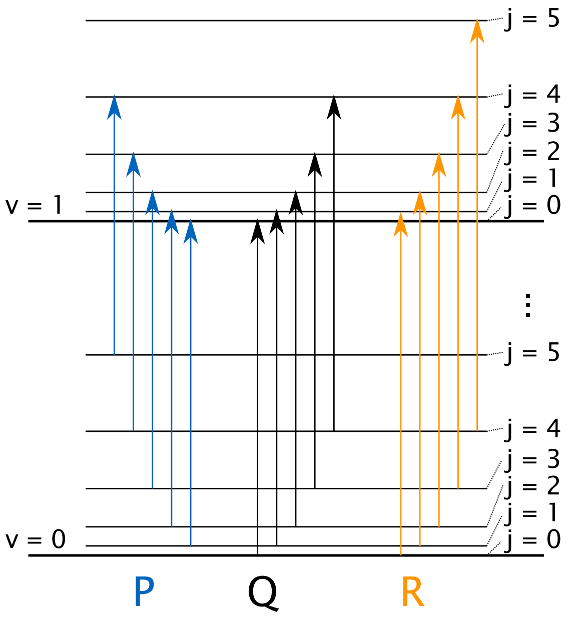


Figure 4: A visualization of exemplary rovibrational transitions of the P, Q, and R-branch. The vibrational quantum numbers (v) and rotational quantum numbers (j) are shown on the y-axis.

Another noteworthy feature of rovibrational spectra is the effect of isotopes. Since isotopes differ only in their weight but not in structure, they show very similar spectra except for a small shift in the peaks of the rotational substructure. [18, Page 534]. When bigger molecules are under investigation, additional effects take place, which makes the interpretation of spectral peaks a lot more difficult. These additional effects stem from a higher interaction of the molecule's atoms and functional groups. The observed effects might even lead to misinterpretation. Some of those will be discussed in the following [20, Page 49 f.]:

- **Coupling** happens between two vibrational modes within the same molecule. This coupling arises due to anharmonicity, leading to deviations from simple harmonic behavior. It includes mode mixing and resonances between similar frequency modes. Especially spatially adjacent groups influence each other's corresponding frequencies.

- **Fermi resonance** can be described as a coupling of a fundamental vibration with an overtone. This leads to a frequency shift and increased intensity of the overtone.

- **Combination bands** occur when one light quantum excites two vibrations simultaneously. The absorbed light quantum has the combined/added energy of the two vibrational states it excites. This further hinders the interpretation since one of those transitions might not be visible in the infrared spectrum due to no change in the dipole moment.

- **Difference bands** signify the transition of an already excited state to a higher energy vibrational state. The energy needed for this transition is the difference between the two states. Difference bands typically have very low intensity.

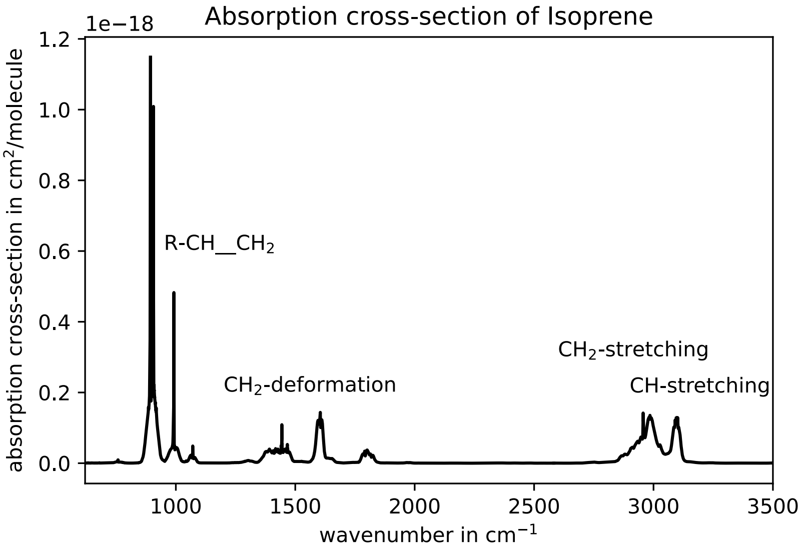


Figure 5: The measured absorption cross-section of Isoprene [28] with annotated transitions. This shows the more complex features of bigger molecules in the fingerprint region.

Even though those effects hinder the interpretation of spectra of bigger molecules, some techniques for spectral interpretation and common markers can be identified. The spectral region from 400-1400 cm$^{-1}$ is often called the fingerprint region due to the great variety of structures, which are hard to attribute to a specific transition. Figure 5 gives an example of transitions and spectral features. A multitude of tables are available to assist in the interpretation of those spectra and to identify specific structures and connect them to a vibrational mode. One is, for example, given in [20, Page 44-59].

## 2.2 Line Intensity

The absorption cross-section of a molecule is connected to the line strength $s$ by:

$$\sigma_A(\tilde{\nu}) = s \cdot f_{LS}(\tilde{\nu}) \tag{11}$$

with $f_{LS}$ as the line shape function. The line shape function will be discussed in more detail in the next Subsection 2.3 and in Chapter 4.1. The parameter $s$ specifies the line strength of a certain transition. For example, $s_{nm}$ specifies the line strength of the transition from state n to state m. It is proportional to the difference of the population densities of the two participating states $\rho_n$ and $\rho_m$ with their corresponding energy levels $E_n$ and $E_m$ as well as the Einstein coefficient for absorption for this specific transition $B_{nm}$, which provides the probability of the transition [18, Page 517]. The full formula is given by:

$$s_{nm} = \frac{\mathrm{h}\tilde{\nu}_{nm}}{\rho \mathrm{c}_0} B_{nm}(\rho_n - \rho_m) \tag{12}$$

Often in literature, the Einstein coefficients of stimulated emission $B'$ and spontaneous emission $A$ are used to express those relations. They can be converted by:

$$B' = B \tag{13}$$

$$A = \left(\frac{8\pi \mathrm{h}\tilde{\nu}^3}{\mathrm{c_0}^3}\right) \cdot B \tag{14}$$

The Einstein coefficient of absorption can be calculated by [18, Page 518]:

$$B = \frac{2\pi^2}{3\epsilon_0 \mathrm{h}^2}|\mu|^2 \tag{15}$$

where $\epsilon_0$ is the vacuum permittivity constant and $\mu$ as the transition dipole moment. This correlates the change of the dipole moment caused by the transition with the height of the observed absorption. Thus, only transitions that show a change in dipole moment $\neq 0$ are visible in the spectrum. The difference in the population states between the two participating states can be calculated:

$$\rho_n - \rho_m = \rho_n \left[1 - \exp\left(-\frac{\mathrm{hc_0}\tilde{\nu}_{nm}}{\mathrm{k}T}\right)\right] \tag{16}$$

using the Boltzmann distribution [20, Page 9]:

$$\frac{\rho_n}{\rho_m} = \exp\left(-\frac{\Delta E}{\mathrm{k}T}\right) = \exp\left(-\frac{\mathrm{hc_0}\tilde{\nu}_{nm}}{\mathrm{k}T}\right) \tag{17}$$

with the temperature $T$ and the Boltzmann constant k. Inserting Equation 16 in 12 we arrive at [29, A1-A3]:

$$s_{nm} = \frac{\mathrm{h}\tilde{\nu}_{nm}}{\rho \mathrm{c_0}} B_{nm}\rho_n \left[1 - \exp\left(-\frac{\mathrm{hc_0}\tilde{\nu}_{nm}}{\mathrm{k}T}\right)\right] \tag{18}$$

The population density of the lower state $\rho_n$ is given by [29, A3]:

$$\rho_n = \rho \frac{\tilde{g}_n}{Q_{int}(T)} \exp\left(-\frac{E_n}{\mathrm{k}T}\right) \tag{19}$$

with $\tilde{g}_n$ as the statistical weight of the lower state $n$ and $Q_{int}(T)$ as the total internal partition function at temperature T. The statistical weight $\tilde{g}_n$ is often called the degree of degeneracy in literature. The total internal partition function is temperature dependent and can be computed by the sum of all internal energy states (rotational, vibrational, electric, etc.) weighted by their statistical weight [29, A4]:

$$Q_{int}(T) = \sum_i \tilde{g}_i \exp\left(-\frac{E_i}{\mathrm{k}T}\right) \tag{20}$$

An in-depth description of how the total internal partition function is computed is given in [18, chap. 20]. Another thorough description as well as tabulated partition functions for many molecules is given in [30], which is also used with the HITRAN database [14]. Since, in most databases, line intensity is

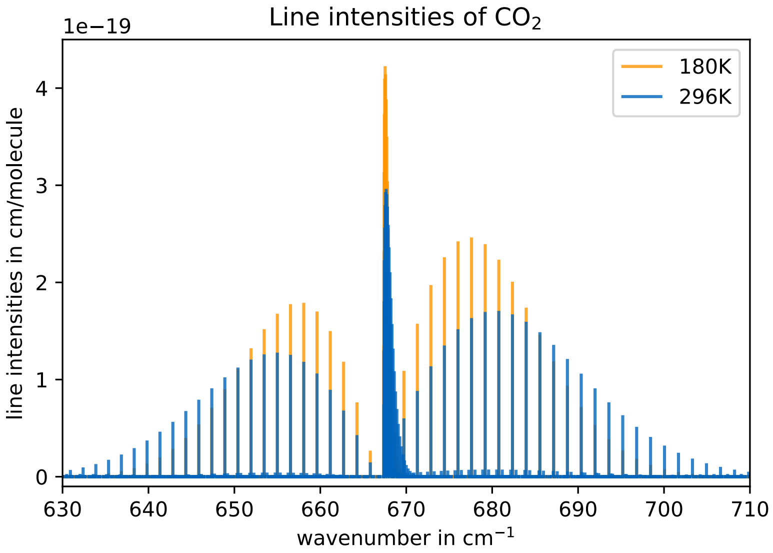


Figure 6: The line intensities of CO$_2$ from HITRAN at 180 and 296 Kelvin. This shows the change in transition probability dependent on the temperature.

one of the parameters used for spectral simulation, the temperature dependency must be removed from the line intensity parameter. To achieve this, HITRAN and other databases provide the line intensity at a reference temperature ($T_{\text{ref}} = 296K$ for the HITRAN database [29]). To get the line intensity at another temperature, a temperature correction has to be applied [29, A11]:

$$s_{nm}(T) = s_{nm}(T_{\text{ref}}) \frac{Q_{int}(T_{\text{ref}})}{Q_{int}(T)} \frac{\exp\left(-\frac{E_n}{\mathrm{k}T}\right)}{\exp\left(-\frac{E_n}{\mathrm{k}T_{\text{ref}}}\right)} \frac{\left[1 - \exp^{\frac{\mathrm{hc}_0 \tilde{\nu}_{nm}}{\mathrm{k}T}}\right]}{\left[1 - \exp^{\frac{\mathrm{hc}_0 \tilde{\nu}_{nm}}{\mathrm{k}T_{\text{ref}}}}\right]} \tag{21}$$

A closer look at computations used for simulating spectral absorption will be provided in Section 4.1. Finally, a simplified visualization of the effect of this temperature-dependent line intensity is provided in Figure 6. Here, the line intensities of CO$_2$ are plotted at two different temperatures. Computations were performed with hapi [27] using line data from the HITRAN database [14]. For the higher temperature, the Q–branch transition from the ground state becomes less pronounced due to a higher probability of molecules being already in an elevated rotational state. The redistribution of the internal partition function is visible through the higher intensities, especially in the outermost parts of the P– and R–branches at higher temperatures.

## 2.3 Line Shape

In the previous sections, transition energies and line intensities were determined. However, in real applications, spectra don't appear as simple, easily separable lines, but each transition line shows a certain line shape of finite width. This shape is described by the normalized function $f_{LS}$:

$$1 = \int_0^\infty f_{LS} \cdot \mathrm{d}\tilde{\nu} \tag{22}$$

This chapter will describe some effects that cause and explain the obtained line shapes. This section will treat the four main factors influencing the line shape: natural broadening, Doppler broadening, collision broadening, and Dicke narrowing. There are additional effects slightly influencing the line shape for which the inclined reader is referred to [31]. Here, the different effects will be described from a physical point of view. In Chapter 4.1, a more practical stance will be adapted as different simulation algorithms, which account for specific effects of broadening, are introduced.

### 2.3.1 Natural Broadening

Even if all other parameters that affect the line shape, like pressure and temperature, would be eliminated by perfect measurement conditions, the spectral line would not be infinitely sharp. The most fundamental effect influencing the line shape is natural broadening. Natural broadening is a quantum mechanical phenomenon linked to Heisenberg's uncertainty principle [32], [22, Page 354]:

$$\Delta E_i \Delta \tau_i \geq \frac{\mathrm{h}}{2\pi} \tag{23}$$

Since the lifetime of molecular state $\tau_i$ can be determined, we can can transform Equation 23:

$$\Delta E_i = \mathrm{hc}_0 \Delta\tilde{\nu}_i = \frac{\mathrm{h}}{2\pi\tau_i} \tag{24}$$

The line shape caused by natural broadening is Lorentzian and determined by the uncertainty of the energies of the upper and lower state as shown in Figure 7. Since the depopulation of the energy levels follows an exponential decay pattern, it results in the Lorentzian line shape $g_{nat}(\tilde{\nu})$ with a Full Width at Half Maximum (FWHM) of $\Delta\tilde{\nu}_{nat} = \frac{1}{2\pi\tau}$. This leads to:

$$g_{nat}(\tilde{\nu}) = \frac{\Delta\tilde{\nu}_{nat}/2\pi}{(\tilde{\nu} - \tilde{\nu}_0)^2 + (\Delta\tilde{\nu}_{nat}/2)^2} \quad \text{with} \quad \Delta\tilde{\nu}_{nat} = \frac{1}{2\pi}\left(\frac{1}{\tau_n} + \frac{1}{\tau_m}\right) \tag{25}$$

after a Fourier transformation to the spectral domain. Thus, the uncertainty of the energy of a molecular state and, therefore, the broadening is inversely proportional to the state's lifetime. Since electronic states have a very short lifetime their natural broadening is in the range of $10^{-4}\mathrm{cm}^{-1}$. Rotational states, on the other hand, show a rather long lifetime leading to natural broadening in the range of $5 \cdot 10^{-15}\mathrm{cm}^{-1}$ [18, Page 521]. Therefore, natural broadening for rovibrational transitions can typically not be observed using standard equipment.

### 2.3.2 Doppler Broadening

The Doppler broadening effect [33] in spectroscopy is very similar to the acoustic Doppler effect well known in daily life. It can be observed when an ambulance passes with the siren seemingly changing frequency after it passes the observer. While the ambulance is driving toward the observer, the acoustic waves are compressed, so their frequency appears higher to the observer. On the other hand, after the

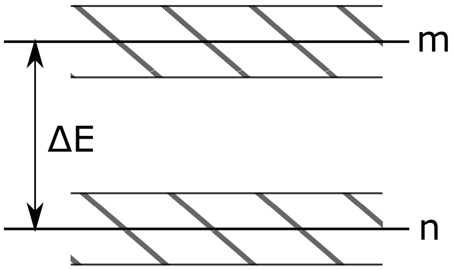


Figure 7: A schematic visualization of the upper (m) and lower (n) state of an exemplary molecular transition with their corresponding uncertainties. Those uncertainties lead to the Lorentzian spectral shape of natural broadening.

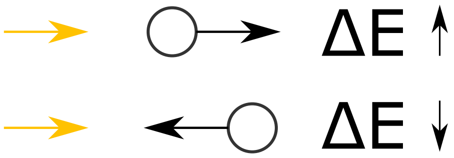


Figure 8: A schematic visualization of the molecule's motion relative to the incident light and its effect on the energy required for a transition. This is commonly known as the Doppler effect.

ambulance has passed the observer, the frequency seems lower since the acoustic waves are stretched due to the movement of the ambulance away from the observer.

Molecules in gas spectroscopy are moving with high velocities in all directions since they show Brownian motion [34], [35]. The distribution of velocities is given by the Maxwell–Boltzmann distribution $p_{MW}$ given by [36], [37]:

$$p_{MW}\left(v\right) = 4\pi\left(\frac{M_i}{2\pi \mathrm{R}T}\right)^{\frac{2}{3}} v^2 \cdot \exp\left(-\frac{M_i v^2}{2\mathrm{R}T}\right) \tag{26}$$

with R as the ideal gas constant and $M_i$ as the molar mass of the molecule $i$. For the spectroscopic Doppler effect, the incident light can be considered the observer, while the molecules show high-velocity motion. Therefore, a shift in the absorbed wavelength can be observed proportional to the molecule's velocity in the direction of the incident light beam as visualized in Figure 8.

This symmetrical effect results in a line shape following a Gaussian distribution. The FWHM $\Delta\tilde{\nu}_{dop}$ of this distribution is given by [18, Page 519]:

$$\Delta\tilde{\nu}_{dop} = \frac{2\tilde{\nu}_0}{\mathrm{c}_0}\sqrt{\frac{2\mathrm{k}T\,\ln\left(2\right)}{M_i}} \tag{27}$$

For simulation purposes the Half Width at Half Maximum (HWHM) $\gamma_D$ is commonly used, which can be simplified to:

$$\gamma_D = \frac{1}{2}\,\Delta\tilde{\nu}_{dop} = \tilde{\nu}_0\sqrt{\frac{T}{M_i}}\,3.581165 \cdot 10^{-7} \tag{28}$$

In general, the Doppler broadening of a molecule increases with increasing temperature, and its effect is more pronounced in lighter molecules. Typical values for the HWHM of Doppler broadening in IR spectroscopy at room temperature are around $10^{-3}\mathrm{cm}^{-1}$ [31, Page 74] and can therefore only be observed with high-resolution measurements at very low pressure.

### 2.3.3 Collision Broadening

Up to now, only the absorbing molecule and its interaction with the incident light were considered for the resulting line shape. However, a gas volume is typically under investigation, containing multiple molecules. Thus, collisions of those molecules have to be taken into account. With increasing pressure, those collisions become more likely. Therefore, all collisional effects are highly dependent on pressure.

For the effect of collisional broadening, often called pressure broadening, we only account for inelastic and dephasing collisions [38]. Inelastic collisions refer to those that lead to energy transfer between the colliding molecules. Dephasing collisions, on the other hand, are defined by a loss of phase coherence within the molecule. The mean time between two collisions is given by $\tau_{col}$, also called the collision lifetime. If we now assume that each of those collisions results in a change in the molecular state, this can, just as for the natural broadening, be linked to Heisenberg's uncertainty principle in Equation 25 leading to:

$$\Delta\tilde{\nu}_{col} = \frac{1}{2\mathrm{c}_0\pi\tau_{col}} \tag{29}$$

Giving the collision FWHM $\Delta\tilde{\nu}_{col}$. The collision lifetime $\tau_{col}$ depends on the relative velocity $\overline{v}_{rel}$ of the two colliding molecules and their collision cross-section given by $\sigma_{col}$. For a system containing only one molecular species, this can be simplified to:

$$\tau_{col}^{-1} = \rho_i \overline{v}_{rel} \sigma_{col} \tag{30}$$

The relative velocity of two molecules is given by [18, Page 876f]:

$$\overline{v}_{rel} = \sqrt{\frac{8\mathrm{k}T}{\pi\mu_{ij}}} \quad \text{with} \quad \mu_{ij} = \frac{m_i m_j}{m_i + m_j} \tag{31}$$

with $\mu_{ij}$ as the reduced mass of the two colliding molecules. The velocity is thus dependent on the interacting molecules' temperature and reduced mass. The collision cross-section, on the other hand, is given by [18, Page 876f]:

$$\sigma_{col} = \pi(r_i + r_j)^2 \tag{32}$$

with $r_i$ and $r_j$ as the radii of molecule i or j, respectively. Thus, the collision lifetime $\tau_{col}$ and, therefore, the broadening parameter $\Delta\tilde{\nu}_{col}$ as well depend on the different components in a gas mixture. Those parameters need to be adapted if the composition of the gas sample changes. When considering a system containing only the analyte species *i* those calculations greatly simplify and Equations 30, 31 and 32 can be combined leading to:

$$\tau_{col}^{-1} = 16r_i^2 p \frac{\pi}{\mathrm{k}Tm_i} \tag{33}$$

Recalling that $\rho_i = N_i \frac{\mathrm{N_A}}{V_m}$ and $V_m = \frac{\mathrm{R}T}{p}$. This reveals the influence of temperature on the broadening coefficient:

$$\Delta\tilde{\nu}_{col,self} \propto \frac{1}{T} \tag{34}$$

Until now, only the state change caused by the collision was considered. The phase shift, on the other hand, results in a shift in the wavelength. This effect stems from interactions between the perturbing molecule and the analyte, which influence the shape of the potential curves [39]. It can cause a negative or positive shift in the peak wavenumber, depending on the strength of deformation for both energy levels. One approximation to deal with this effect in line shape modeling is to add an imaginary (frequency shifting) part to the correlation function $C(t)$ [31], [40, Page 77]:

$$C(t) = e^{-\left(1+i\Delta_{shift}\right)t} \tag{35}$$

With the frequency shift given by $\Delta_{shift}$. The Fourier transform of the correlation function leads to the line shape of the transition in question, and a frequency shift here results in a shift in wavenumbers. Combining those two effects leads to an unnormalized line shape function $g_{col}(\tilde{\nu})$:

$$g_{col}(\tilde{\nu}) = \frac{\Delta\tilde{\nu}_{col}/2\pi}{\left(\tilde{\nu} - \tilde{\nu}_0 - \Delta_{shift}/2\right) + \left(\Delta\tilde{\nu}_{col}/2\right)^2} \tag{36}$$

This function is also known as the Lorentz profile and will be described further in Chapter 4.1.

### 2.3.4 Dicke Narrowing

The last effect that will be presented is the so-called Dicke narrowing [41]. Dicke narrowing characterizes one of the effects of velocity-changing collisions. If we consider an analyte molecule with a certain velocity $v$ along the same axis as the light beam, this molecule is bound to collide with other molecules. If those collisions have a velocity-changing effect on the analyte molecule, this effect will, on average, decrease the velocity of the molecule in question. With the decrease of the average velocity of the molecule follows a narrowing of the Doppler broadening. This effect is independent of the direction of the molecule's velocity. The HWHM $\Delta\tilde{\nu}_{dicke}$ of the dicke narrowing is inversely proportional to the pressure [31, Page 78]:

$$\Delta\tilde{\nu}_{dicke} \propto \frac{1}{p^2} \tag{37}$$

The effect of Dicke narrowing becomes only applicable at high pressures and is, therefore, barely considered during laboratory spectroscopy measurements at atmospheric conditions.

The effects presented in this section outline the most prominent influences that change the line shape of an absorption. It is not comprehensive, and the inclined reader is referred to [31] for additional reading. Nevertheless, the fundamentals of the physics of absorption and the most critical influences on the line shape and height have been described. This knowledge provides a basis for the simulation of spectra in Chapter 4 and the deep learning system for spectral adaption presented in Section 6. The next chapter presents the spectroscopic methods that exploit the physical effect of absorption to analyze a gas mixture.

# 3 Spectroscopic Methods

The physical processes of molecular absorption can be utilized to analyze gas mixtures in multiple ways. This chapter provides an overview of optical gas absorption spectroscopy methods, including an in-depth description of the photoacoustic measurement principle. This measurement principle will later form the basis for generating synthetic photoacoustic spectra.

Most optical gas spectroscopy methods are based on the Lambert-Beer law presented in the first section of this chapter. The application fields of optical gas sensing are extensive, covering health care, energy, air pollution monitoring, industry quality monitoring, and many more [42]. Similarly, the applications span a wide range of different requirements. Therefore, various techniques have been developed to satisfy those demands. Here, a small subset of all possible sensing methods will be presented, which are the most common absorption spectroscopic methods. Additional focus will be placed on photoacoustic trace gas analysis.

First, different sensing systems focused on quantitative absorption detection of an analyte are outlined. Afterward, an introduction to the photoacoustic measurement method is presented. This includes the signal generation process, as well as different measurement setups and modulation techniques. Finally, the non-spectral effects complicating Photoacoustic Spectroscopy (PAS) are outlined.

## 3.1 Lambert-Beer

Up to now, only the physical absorption process has been qualitatively discussed. Quantitative IR and UV / VIS spectroscopy relies on the measured quantities transmittance $\tilde{T}(\tilde{\nu})$ and absorbance $\tilde{A}(\tilde{\nu})$ of a sample at a certain wavenumber. The relation between absorbance and transmittance is given by:

$$\tilde{T}(\tilde{\nu}) = \frac{I_1(\tilde{\nu})}{I_0(\tilde{\nu})} \quad \text{and} \quad \tilde{A}(\tilde{\nu}) = -\ln \tilde{T}(\tilde{\nu}) \tag{38}$$

with $I_0(\tilde{\nu})$ and $I_1(\tilde{\nu})$ being the incident and transmitted light intensity. Attention must be paid to the fact that absorption $A'(\tilde{\nu})$, which is in this work used as a qualitative physical term, often refers to the amount of light that is absorbed by the sample and must not be confused with absorbance $\tilde{A}(\tilde{\nu})$:

$$A'(\tilde{\nu}) = \frac{I_0(\tilde{\nu}) - I_1(\tilde{\nu})}{I_0(\tilde{\nu})} = 1 - \frac{I_1(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{39}$$

The transmitted light intensity can also be expressed by:

$$I_1(\tilde{\nu}) = I_0(\tilde{\nu}) e^{-\tilde{A}(\tilde{\nu})} \tag{40}$$

which is commonly known as the Lambert-Beer Equation [43], [44]. The absorbance can be replaced by:

$$\tilde{A}(\tilde{\nu}) = \epsilon(\tilde{\nu}) c_i l_{opt} \tag{41}$$

where $c_i$ is the concentration of the analyte $i$, $l_{opt}$ is the optical pathlength and $\epsilon(\tilde{\nu})$ is the extinction coefficient at that wavenumber. This shows that the absorbance is directly proportional to the optical path–length and the concentration. Equation 41 holds if only one analyte and no other components absorb at the corresponding wavenumber; otherwise, it needs to be extended. For gaseous samples Equation 41 is normally rewritten to:

$$I_{1,naperian}(\tilde{\nu}) = I_0(\tilde{\nu})e^{-\alpha(\tilde{\nu})l_{opt}} \tag{42}$$

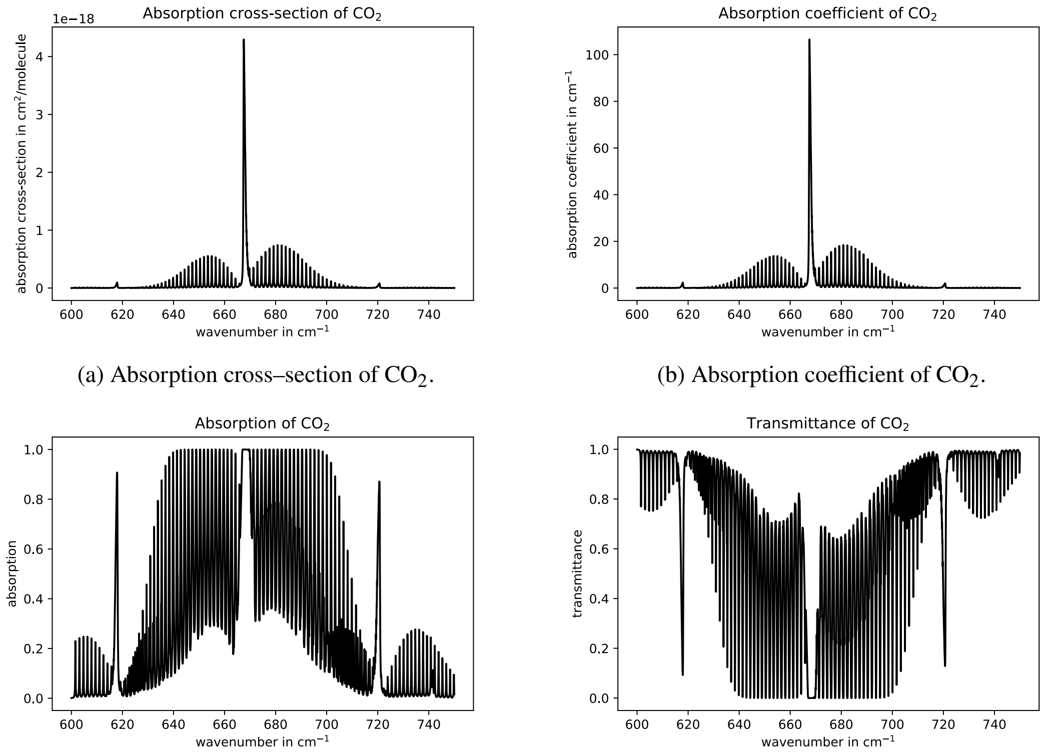
Using $\alpha(\tilde{\nu})$ the absorption coefficient, which replaces $\epsilon(\tilde{\nu})c_i$. Besides, absorbance can be expressed in Naperian (base-e, 42) or logarithmic base (base-10, 43). The Naperian notation has historical reasons and has been retained in some databases. The logarithmic version can be computed as:

$$I_{1,log}(\tilde{\nu}) = I_0(\tilde{\nu})10^{-\alpha(\tilde{\nu})l_{opt}} \tag{43}$$

This has to be kept in mind when considering spectra from different databases as explained more deeply in Chapter 4.2. Another commonly used quantity is the absorption cross-section $\sigma_A(\tilde{\nu})$, which can be converted to the absorption coefficient by:

$$\alpha(\tilde{\nu}) = \sigma_A(\tilde{\nu})\rho_i \quad \text{with} \quad \rho_i = N_i \frac{\mathrm{N_A}}{V_m} \tag{44}$$

Where $\rho_i$ is the particle density of the analyte, $N_i$ is the volume ratio of the analyte $i$ and $\mathrm{N_A}$ is the Avogadro constant and $V_m$ is the molar volume. Since the particle density $\rho_i$ depends on environmental parameters such as pressure and temperature, the absorption cross-section can be regarded as a version of the absorption coefficient independent of ambient pressure. Absorption cross–section (9a), the absorption coefficient (9b) as well as the absorption (9c) and transmittance (9d) of the molecule $CO_2$ are shown in Figure 9. Those plots were created using the HITRAN [30] line–database and the Voigt [26] simulation algorithm in hapi [27]. HITRAN and hapi use Naperian notation. In the transmittance and absorption spectra (9c, 9d), the effect of saturation can be seen. Here all light is absorbed $A'(\tilde{\nu}) = 1$ and none is transmitted $\tilde{T}(\tilde{\nu}) = 0$. This is an undesirable effect for spectroscopy, which can be prevented by decreasing the path–length or smaller concentrations.



(a) Absorption cross–section of $CO_2$.

(b) Absorption coefficient of $CO_2$.

(c) Absorption spectrum of 100 % $CO_2$ for a path–length of 1 cm.

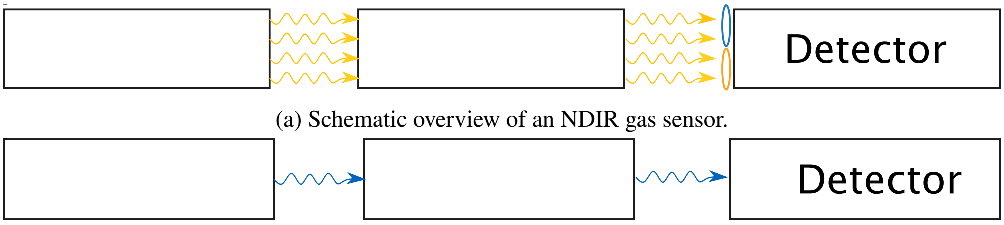
(d) Transmission spectrum of 100 % $CO_2$ for a path–length of 1 cm.

Figure 9: Absorption cross-section (a), absorption coefficient (b), absorption spectrum (c), and transmittance spectrum (d) of $CO_2$ simulated using the HITRAN database and a Voigt profile. For the absorption (c) and transmittance spectrum (d) the saturation effect can be seen.

## 3.2 Non-Dispersive Infrared Sensors

The most widespread and most straightforward technique in optical gas sensing is Non–dispersive Infrared (NDIR), which was first applied in 1943 [45]. In typical NDIR sensors, a broadband emitter in the NDIR wavelength regime is used, which passes through the gas volume. Two different filters are placed in front of the detector, one covering the absorbing region and one a non-absorbing region of the analyte. Those two signals' differences can now be used to determine the analyte concentration. This technique is visualized schematically in Figure 10a. This approach is often extended to include a dual beam approach or rotating filters [42]. It can be used for multi-component sensing by introducing multiple filters.

Tunable Diode Laser Absorption Spectroscopy (TDLAS) is conceptually similar to NDIR but uses a laser as an emitter, thus reaching a very narrow emission peak, which can be tuned to an analyte absorption line. This concept is visualized in Figure 10b. It is often coupled with a two-beam approach employing a reference channel filled with a reference sample. NDIR and TDLAS spectroscopy only use small spectral points; thus, they do not rely on the line shape of the spectra or multiple absorption lines. Nevertheless, precise knowledge of the peak positions is needed to select the wavelengths in question. Spectral interference with other substances present in the sample in one of those regions poses a significant problem for those sensors. On the other hand, this concept lends itself well to miniaturization and the production of specialized, possibly cheap optical gas sensors. The two approaches NDIR and TDLAS measure the transmission, which can be converted to the absorption of a sample. To increase the sensitivity of those methods, wavelength modulation is often employed. An introduction to wavelength modulation is provided to the inclined reader in [19, p. 64 ff]. Nevertheless, the sensitivity of NDIR and TDLAS is limited by the optical path length and the detector quality.



(a) Schematic overview of an NDIR gas sensor.

(b) Schematic overview of a TDLAS gas sensor.

Figure 10: Schematic overview of the NDIR (a) and TDLAS (b) gas sensing approaches. For the NDIR (a) sensor, broadband light (yellow waves) passes through the probe and two different filters (blue, orange) before the detector while at the TDLAS (b) sensor laser light (blue wave) covering only a small wavelength region passes through the probe.

## 3.3 Fourier Transform Infrared Spectroscopy

There are two main approaches concerning spectrometers, which cover a broad part of the spectrum. They differ in acquisition type. Both methods employ a broadband light source as their emitter. The scanning approach uses a monochromator such as a prism or an optical grating to differentiate the wavelengths. This monochromator is moved over time and performs a full scan of the spectrum. One of the main disadvantages of this approach is the long duration of one measurement, up to a few seconds or even minutes [20, p. 43]. On the other hand, it is very well suited for studying pulsed signals.

A Fourier Transform Infrared (FTIR) spectrometer is the other, more common approach. FTIRs employ a Michelson interferometer to differentiate the wavelength signal [46]. A Michelson interferometer splits the original beam into two parts through a beam splitter. Both beams are reflected back on mirrors and combined again before they reach the detector. Due to the differing path lengths, the two beams now interfere. One mirror is kept at a constant distance, and the other is moving; thus, the relative path length between the two beams is altered. This leads to the two beams being either in or out of phase, resulting in an interferogram. A Fourier transformation can be employed to transfer this interferogram from the time domain to the frequency domain, hence providing the transmission spectrum. This provides multiple advantages over the traditional scanning approach, such as higher measurement speed and a better signal-to-noise ratio [20, p. 44]. FTIR are commonly used for (gas) analysis in many laboratories and provide a baseline for most sensor developments. Even though they allow for sensitive and selective quantification of gases, they are often bulky, expensive, and require trained staff, which limits their application in consumer technology.

## 3.4 Photoacoustic Spectroscopy

When handling trace gases at very low concentrations, the transmission shows a very high level. The incident light intensity only changes slightly due to the absorption of the analyte. This creates the need for expensive high-resolution detectors and long optical paths to allow for the induced signal alteration to be detected. A technique to circumvent this problem is PAS. PAS can be considered a direct measurement technique, using the pressure changes caused by light absorption and subsequent non-radiative relaxation processes. It was first discovered in 1880 [47] and first employed in gas analysis in 1938 [48].

A laser serves as an emitter for most current PAS setups. The laser is modulated or manually chopped to generate a modulated signal. During the on phase, the analyte absorbs the light and reaches higher rovibrational states. During the off phase, relaxation occurs. If this relaxation is non-radiative, this leads to heat production, which can be detected as a change in the pressure. Since on and off phases occur periodically, this pressure signal can be measured as a microphone's soundwave of the modulation frequency. Its amplitude is directly proportional to the absorption signal of the probe. Hence, without any absorption, no signal is generated, which renders this measurement method a so-called direct method. A schematic overview is given in Figure 11. While tackling the problem of the high signal level, PAS itself has some challenges of its own, including that only non-radiative relaxation processes can be detected, and additional non-spectral interferences have to be accounted for. Modulation methods to increase sensitivity are available, similar to absorption spectroscopy.

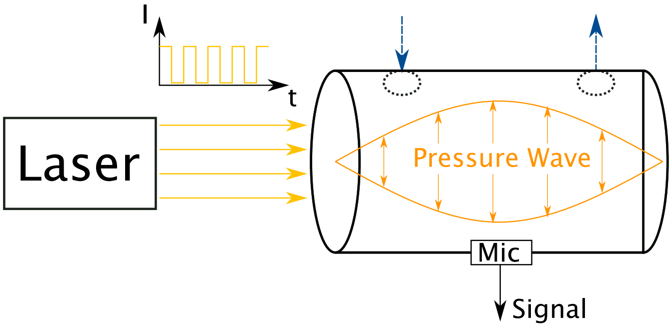


Figure 11: Schematic concept of a photoacoustic gas sensor. The modulated laser beam (yellow) is generated by a rectangle shaped driving current. It generates a standing pressure wave within the tube shaped cell filled with the gas sample. This pressure wave is measured using a microphone.

### 3.4.1 Signal Generation

The photoacoustic signal is generated in multiple steps, as visualized in Figure 12. Since the absorption process has already been described in the previous sections, we focus on heat production and acoustic wave generation, which is specific to PAS. First, we focus on the heat production rate $\dot{H}(t)$, which can be computed as [49]:

$$\dot{H}(t) = \dot{H}_0 e^{i(\omega t - \phi)} \tag{45}$$

with

$$\dot{H}_0 = \rho\sigma(\tilde{\nu}) I_0 \epsilon_{relax} \tag{46}$$

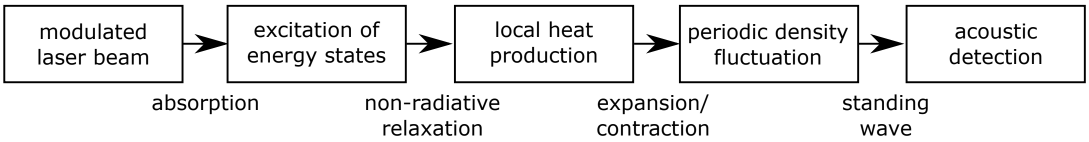


Figure 12: Schematic of the photoacoustic signal generation principle. (adapted from [49])

The heat production rate modulates over time with the angular frequency of the laser modulation $\omega$ and the phase lag between the modulated laser beam and local heat input $\phi$. The heat production rate is often also called power density. Its amplitude $\dot{H}_0$ depends on the volume particle density $\rho$, the absorption cross-section $\sigma$ at the laser output wavenumber $\tilde{\nu}$ as well as the incident light intensity $I_0$ and the efficiency of the molecules relaxation $\epsilon_{relax}$. In this equation, the laser output is simplified to be at precisely one wavenumber. The absorption cross-section and the relaxation efficiency need to be computed for the exact combination of molecules in a mixture.

The efficiency of the molecule's relaxation of an individual energy state $v$ can be defined as

$$\epsilon_{relax,v} = \frac{1}{\sqrt{1 + (\omega\tau_v)^2}} \tag{47}$$

where $\tau_v$ is the non-radiative lifetime constant of $v$ and $\omega$ is the angular frequency of laser modulation. If 100% of the excited states can return to their ground state within one laser cycle, the relaxation efficiency becomes $\epsilon_{relax} = 1$. The total lifetime of an exited state $\tau_v$ is the reciprocal sum of the radiative $\tau_r$ and non-radiative time $\tau_n$ constants. The total lifetime of an excited state $\tau_v$ corresponds to the time until $\frac{1}{e}$ of the corresponding state decays.

$$\tau_v^{-1} = \tau_n^{-1} + \tau_r^{-1} \tag{48}$$

The radiative lifetime corresponds to a radiative transition, including a generated photon. This radiative lifetime for excited vibrational and rotational states is relatively long (0.01 - 1 s) [50]. Therefore, in a PAS setting where the laser signal is modulated at a higher frequency than radiative relaxation is possible, it can often be neglected. This topic is covered in more detail in Section 3.4.4.

The acoustic wave generation is governed by multiple physical laws, including the momentum conservation law, mass conservation law, energy conservation law, classical fluid mechanics, classical thermodynamics, and the thermodynamic equation of state [51]. Its complete derivation is beyond the scope of this thesis, and the inclined reader is referred to [19, Page 80-84]. The solution of the non-damped inhomogeneous equation links the frequency-dependent heat production rate with the generated acoustic wave by [52]

$$\left(\Delta + \frac{\omega^2}{c_s^2}\right) p_a\left(\vec{r}, \omega\right) = \left(\frac{\gamma - 1}{c_s^2}\right) i\omega\dot{H}\left(\vec{r}, \omega\right) \tag{49}$$

with

$$p_a\left(\vec{r}, t\right) = \int \mathrm{d}\omega p_a\left(\vec{r}\omega\right) e^{-i\omega t} \tag{50}$$

and

$$\dot{H}\left(\vec{r}, t\right) = \int \mathrm{d}\omega\dot{H}\left(\vec{r}\omega\right) e^{-i\omega t} \tag{51}$$

Where $c_s$ is the speed of sound, $\omega$ the angular frequency of laser modulation, $\Delta$ the Laplace operator, $\vec{r}$ the three-dimensional position vector, $p_a$ the absorption-induced sound pressure, and $\gamma$ the heat capacity ratio. Using a typical tube-shaped resonator at the resonator eigenfrequencies $\omega_j$ of the $j^{th}$ normal mode, the final photoacoustic signal can be described with

$$p_a\left(\vec{r}, \omega_j\right) = C_{cell}\epsilon_{relax}\rho\sigma(\tilde{\nu})P_o e^{-\mu_j} p_j\left(\vec{r}\right) \tag{52}$$

with the cell constant $C_{cell}$

$$C_{cell} = (\gamma - 1)\frac{Q_j}{\omega_j}\frac{L_R}{V_R}p_j \tag{53}$$

where $P_0$ is the incident optical power, and $\mu_j$ the reciprocal light-to-sound coupling factor of the $j^{th}$ mode. The pressure distribution $p_j\left(\vec{r}\right)$ is the only factor depending on spatial coordinates. In the case of the first longitudinal mode excitation, this simplifies to a sinusoidal half-wave from 0 to 1 [53]. The cell constant depends on the mode's $Q_j$ factor, the length and volume of the cell $L_R$ and $V_R$, and the normalization coefficient $p_j$. Again, in the first longitudinal excitation mode, this normalization factor can be simplified to $p_j = \sqrt{2}$ [52], [53]. For a more in-depth explanation and additional resonator geometries, the reader is referred to [19, chap. 2.3].

The acquired photoacoustic signal is a complex signal consisting of the pressure magnitude and phase in relation to the laser modulation. While the signal amplitude is used for most applications, the phase also holds information, especially about relaxation pathways. Finally, this acoustically resonator-enhanced pressure signal is acquired by a pressure transducer, often a microphone. The method of detection of the signal greatly depends on the selection of the measurement setup.

### 3.4.2 Measurement Setups

A broad range of photoacoustic measurement setups exist that are specialized for different purposes. A full description of the wide variety of different setups is beyond the scope of this thesis. The two setups used during this thesis will be explained in this chapter, while resources on other principles and setups are provided. Photoacoustic measurement setups can differ in their use of different light sources and their usage, as well as the photoacoustic cell design and choice of signal transducer. Weigl et al. [49] provide a good overview of the different techniques, focusing on the design of a breath analysis PAS sensor.

While broad-range light sources are used for FTIR spectroscopy, in PAS, lasers and diodes are more common. The criteria for the choice of light source selection include the optical power of the light source, the output wavenumber, the precision of the output wavenumber, beam configuration, and cost. While diodes are typically relatively cheap, mainly if they are also applied in other areas, for example, telecommunication, they do not allow tuning of the output wavenumber and usually do not provide a focused beam profile. Recently, Interband Cascade Laser (ICL) and Quantum Cascade Lasers (QCLs) have been increasingly employed for multi-component detection due to their tunability of a few wavenumbers and the higher optical power. In the research context, sometimes Optical Parametric Oscillators (OPOs) have been applied, which are a laboratory instrument with a very broad tuning range, far beyond the capabilities of ICLs and QCLs. The high cost of OPOs discourages their use in actual sensor products. Other light sources, such as $CO_2$–lasers, are also possible but uncommon. A QCL and an ICL have been employed in this work.

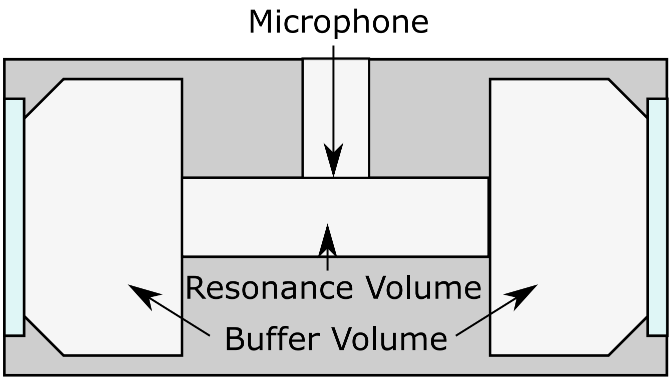


Figure 13: Schematic of an H-shaped resonant photoacoustic cell design as employed in the QCL measurement example. The windows are shown in light blue, while the gas volumes remain white. Two buffer volumes enclose the resonance volume in the middle that holds the standing pressure wave. The microphone is placed at the middle where the highest amplitude change is expected.

One of the two measurement setups used within this thesis is an H-shaped resonant cell, with a microphone placed in the middle of the tube as visualized in Figure 13. Two buffer volumes enclose the resonance volume in the middle. A simple Micro Electro Mechanical System (MEMS) microphone is placed in the middle of the resonance volume, detecting the pressure change. The acoustic resonance gained by this setup has been described in the previous section and can be further increased by using a Lock-in Amplifier (LIA) as the modulation frequency is known.

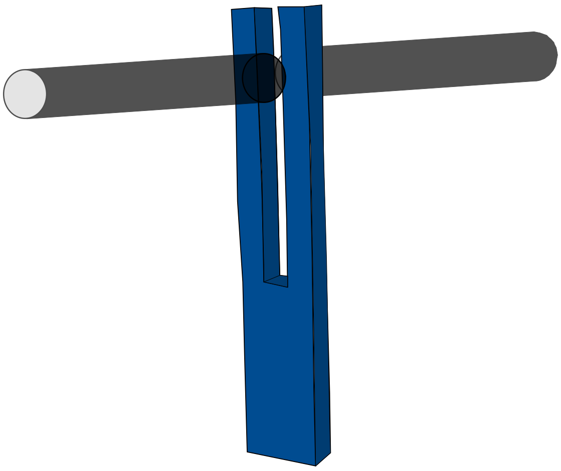


Figure 14: QEPAS photoacoustic prong measurement setup as employed in the ICL measurement example. The laser is focused between the two prongs to excite their resonance frequency. The tube resonators enhance the signal.

The second measurement setup is a Quartz-Enhanced Photoacoustic Spectroscopy (QEPAS) setup as visualized in Figure 14. The QEPAS setup achieves the signal amplification by exciting the quartz tuning forks' resonance, which replaces the microphone in this setup. Additional acoustic resonance can be gained by the tube resonator placed in close proximity to the prongs. Quartz Tuning Forks (QTFs) require elaborate beam focusing of the light source through the prongs. QTFs also provide better ambient noise immunity than standard microphones [54]. On the other hand, the QTF double-resonant systems are more prone to detuning, which increases the difficulty of calibration [49], [55].

### 3.4.3 Modulation Techniques

To generate a photoacoustic signal, the light has to be modulated. Two distinct modulation techniques are commonly used in PAS. Those are referred to as Amplitude Modulated Photoacoustic Spectroscopy (am-PAS) and Wavelength Modulated Photoacoustic Spectroscopy (wm-PAS). In amplitude modulation, the light is periodically turned on and off. Often, this chopping can be achieved by modulation of the operating current of the light source. An alternative is a mechanical chopper [56]–[58]. The duty cycle of the modulation is typically at around 50 % but can be far lower for pulsed signal generation [59].

The wavelength modulation is illustrated schematically in Figure 15. The light source driver supplies a constant offset current $I_{\text{Offset}}$. The light source, therefore, generates an output at the center wavelength $\lambda_c$. The output is modulated around this center wavelength. To achieve this, typically, a small sinusoidal modulation current $I_{\text{mod}}$ on top of the offset current is applied. Thus, the laser output wavelength is modulated, and when intercepting with a slope in the absorption signal, an acoustic PAS signal is generated. This $2f$-signal increases in height as the slope of the absorption increases. Thus, not the absorption signal itself is measured but its derivative.

When WM-PAS is applied as an on-peak measurement technique, the modulation depth is optimized to correspond to the FWHM of the peak in question to increase the PAS signal. For spectral acquisition, the offset current is slowly ramped during the measurement, resulting in multiple WM-PAS measurement points. In this case, the modulation depth is a compromise between signal intensity and slope correctness. While the $2f$-signal generation of this technique brings several advantages, i.e., the constant absorption background does not contribute to the signal, it also makes the signal more sensitive to changes in peak shape, often resulting from pressure changes.

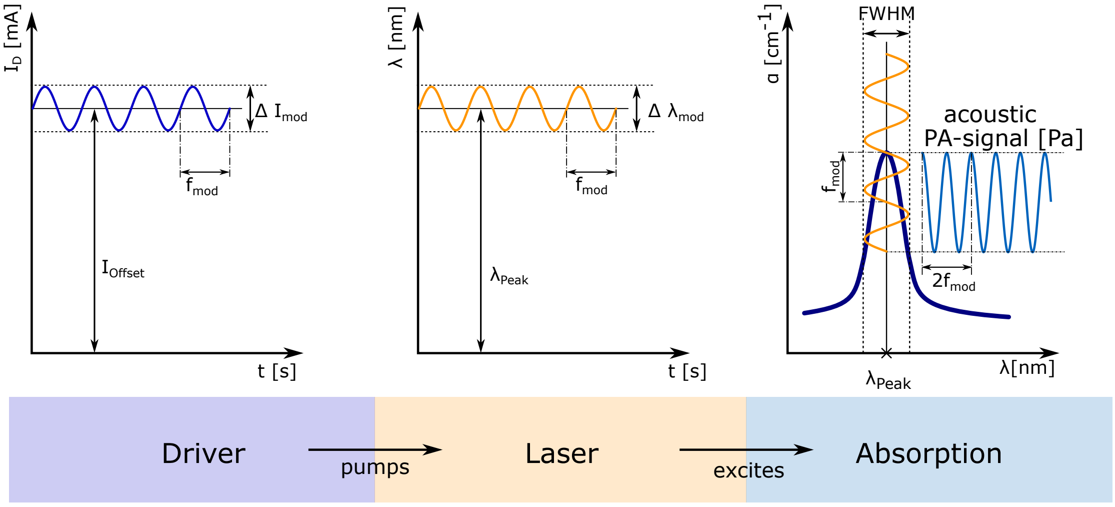


Figure 15: The wavelength modulation principle in photoacoustic spectroscopy. The laser driver generates a current signal modulated around a center current $I_{Offset}$. This signal pumps the laser to emit light oscillating around a center wavelength $\lambda_{peak}$. The rightmost plot shows the absorption coefficient (dark blue) of the transition over the wavelength. The incident light generated by the laser is indicated in orange and oscillates around the center wavelength. This oscillating signal generates the light blue acoustic pressure signal. (Adapted from [49])

### 3.4.4 Non-Spectral Effects

Even though PAS can achieve great sensitivity for trace gas spectroscopy as it is a direct measurement technique, it comes with some disadvantages. This includes additional non-spectral effects that differentiate the photoacoustic signal from traditional absorption spectra. Those have to be considered, especially in complex gas mixtures. The most critical non-spectral effects include:

- Non-radiative molecular relaxation
- Acoustic attenuation
- Changes in heat capacity ratio

### Non-radiative molecular relaxation

Non-radiative molecular relaxation relates to the parameter $\epsilon_{relax}$ introduced in Equation 47. It simplifies to one if all molecules excited by the absorbed light relax non-radiatively during one laser cycle. Those relaxation processes include collisional relaxation processes. When the excited molecule collides with another molecule, it can transfer all (VT relaxation) or part of its vibrational energy (VV relaxation) as translational energy. VV relaxation can happen as an inter- or intramolecular energy transfer, so the translational energy is transferred to the collision partner or onto itself. Those processes are visualized in Figure 16.

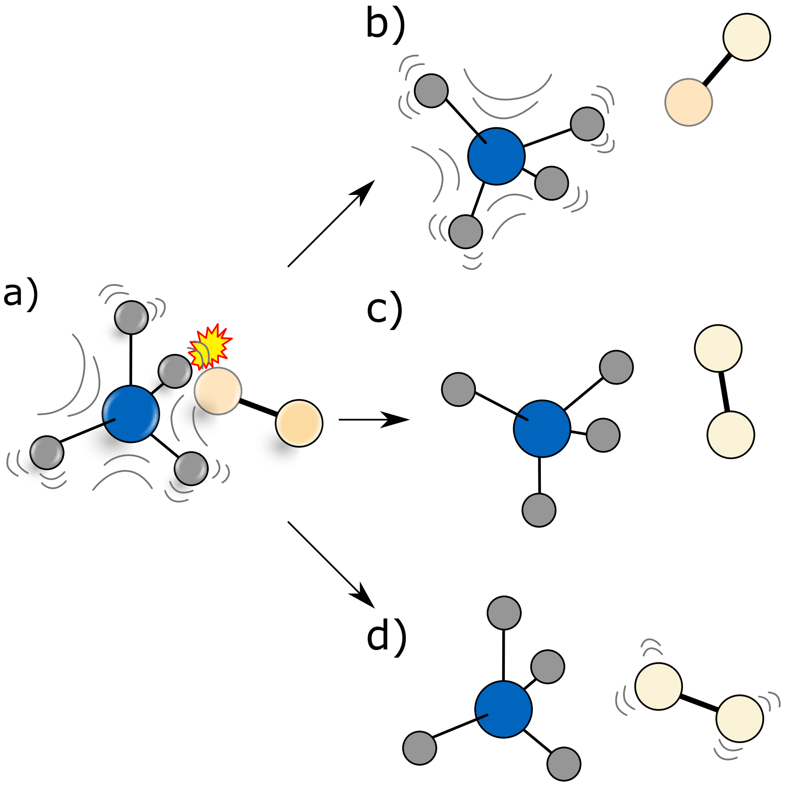


Figure 16: Schematic representation of the non-radiative relaxation processes. After a collision with another molecule (a), intramolecular energy transfer (b), a VT relaxation (c), or a VV relaxation (d) can take place. (adapted from [49])

With more complex gas mixtures, the number of possible intermolecular transitions increases significantly, rendering the computation of the radiative lifetime more complex. To compute the radiative lifetime correctly, all possible transitions must be included. The exact efficiency of the relaxation also depends on the gas composition and pressure, as the number of collisional partners impacts the probability of a particular transition happening. Those thoughts have long been neglected and simplified in PAS [16]. Recently, analytical approaches to deal with those relaxational effects in changing gas compositions have been reported that show good agreement with measured data [53], [55], [60], [61]. An algorithm to compute the collision-based non-radiative efficiency has been introduced by Müller et al. [16]. It has been shown exemplarily on a NIR methane transition. Even though this algorithm and the previous analytical solutions allow a simulation of the relaxational-induced signal alterations, many transitional parameters are needed for each composition and transition, which are often hard to obtain.

**Acoustic attenuation**

From Equation 49, the dependence of the standing wave in the acoustic resonator on the speed of sound $c_s$ becomes evident. When the gas matrix changes significantly, the speed of sound of the medium also changes, leading to shifts in the resonance frequency of the system. In this case, the resonance frequency needs to be adapted; otherwise, substantial deviations from the calibrated value can be observed [62].

**Heat capacity ratio**

The cell constant of the resonator depends on the heat capacity ratio $\gamma$ of the gas matrix corresponding to the ratio of the specific isobar and isochore heat capacities, $C_p$ and $C_V$. This heat capacity ratio of gaseous media can generally be expressed as

$$\gamma = \frac{C_p}{C_V} = 1 + \frac{2}{\sum f} \tag{54}$$

where $\sum f$ is the sum over all Degree of Freedoms (DOFs). Those DOFs include linear motion and the molecules' rotational and vibrational states. At lower temperatures and faster acoustic frequencies, no longer all vibrational DOFs can be considered to contribute to sound propagation, leading to changes in $\gamma$. A more detailed description and derivation can be found in [49, Chap. 4.3.1.] and [63].

This chapter provides an overview of the spectroscopic methods in optical gas spectroscopy, focusing on photoacoustic signal generation. This knowledge is applied in Chapter 7 to generate synthetic photoacoustic spectra. Another important part of synthetic photoacoustic spectra is the underlying absorption spectra, whose computation will be presented in the following chapter.

# 4 Absorption Modeling

The physical processes described in Chapter 2 can be used to simulate synthetic absorption spectra. The common toolchains for their computation are described in this chapter. The simulated absorption spectra form the basis of the synthetic photoacoustic spectra presented in Chapter 7 and the training data for the adaption deep learning model presented in Chapter 6.

Leveraging the knowledge of the physical effects governing molecular absorption, modeling the absorption spectra and their characteristics is possible. With those methods, an expected spectrum of the absorption of a specific gas composition can be obtained. The line data-based simulation can be compared to and evaluated against an actual, measured spectrum. This kind of simulation allows for determining the composition of a gas mixture.

In the following section, the modeling of absorption spectra will be introduced. After introducing two line shape models, the easy-to-determine Voigt and very precise Hartman-Tran model, the different line databases currently available online are presented and compared. Next, an introduction to the applied simulation code provided by hapi [27] is given. Finally, the limitations of the line-based modeling approach are discussed. The available cross-section databases and additional methods to simulate absorption spectra are presented as alternatives.

## 4.1 Line Shape Models

Various line shape models exist, incorporating various molecular effects and requiring a different amount of compute and input parameters. Therefore, researchers and sensor developers must select a line shape model applicable to their use case. One point of consideration are the ambient parameters at which the measurement will be conducted, as different effects dominate the line shape at different ambient configurations. For example, the Doppler effect shows at very low pressures, while Dicke narrowing only becomes notable at high pressures, rendering specific effects important for different ambient configurations. Another parameter to consider is the availability of line shape parameters. Simpler profiles like the Voigt model, which will be explained in Section 4.1.1 only require a few parameters, while more intricate models, as the Hartman-Tran model presented in Section 4.1.2 require many additional line shape parameters which might not be available for many transitions or molecules. Last, computation time and the needed wavenumber resolution must be considered for line shape selection. One spectrum often consists of many thousands of transitions, and line-shape algorithms are usually not yet fully parallelized. In this subsection, the most common line-shape algorithms are presented. The section is intended to provide an overview of available line shapes but is by no means complete. A more exhaustive overview can be found in [31]. All algorithmic descriptions used here are based on their implementation in hapi (Version 1.1.0.9.7) [27].

### 4.1.1 Voigt Model

The widely used Voigt profile [26] combines the dephasing collision broadening presented in Section 2.3.3 and the temperature induced Doppler broadening presented in Section 2.3.2. It is, therefore, applicable, especially in the medium pressure range at normal temperatures, which corresponds to the ambient configuration on Earth. It also provides the basis for many more intricate line-shape models [31, p. 79ff].

The Voigt profile is the Laplace transform of the product of the two time-dependent dipole autocorrelation functions of the Doppler and Lorentz profile normalized to area 1 [31, p. 79]:

$$I_V(x, y) = \int_{-\infty}^{+\infty} I_D(x')\, I_L(\tilde{x} - x', y)\, dx' \tag{55}$$

Where $I_V$ is the voigt profile, $I_D$ the Doppler profile and $I_L$ the Lorentz profile. The variables $x, \tilde{x}$ and $y$ are dimensionless parameters introduced by [64], [65], to ease computation. They are defined as follows:

$$x = (\tilde{\nu} - \tilde{\nu}_0)\frac{\sqrt{\ln(2)}}{\gamma_D} \tag{56}$$

$$y = \frac{\sqrt{\ln(2)}}{\gamma_D}\gamma_0 \tag{57}$$

$$\tilde{x} = x - \frac{\sqrt{\ln(2)}}{\gamma_D}\Delta_{shift} \tag{58}$$

Where $\gamma_0$ is the HWHM of the collisional broadening which corresponds to $\frac{1}{2}\Delta\tilde{\nu}_col$ described in Equation 29 and $\gamma_D$ the HWHM of the Doppler broadening. Using those parameters, the Lorentz and Doppler profile functions can be defined as follows [31, p. 79ff]:

$$I_D(x) = \frac{1}{\sqrt{\pi}}\mathrm{e}^{-x^2} \tag{59}$$

$$I_L(\tilde{x}, y) = \frac{1}{\sqrt{\pi}}\frac{y}{\tilde{x}^2 + y^2} \tag{60}$$

Hence, the function for the Voigt profile from Equation 55 can be rewritten as:

$$I_V(x, y) = \frac{1}{\pi}\frac{y}{\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{e}^{-x'^2}}{(\tilde{x} - x')^2 + y^2}dx' = \frac{1}{\pi}K(\tilde{x}, y) \tag{61}$$

where $K(\tilde{x}, y)$ is the Voigt function which is normalized to area $\sqrt{\pi}$. An extensive description of the Voigt function can be found in [66]. It is not analytically solvable. The Voigt function corresponds to the real part of the complex probability function [67, p. 298 ff]:

$$w(x, y) = \frac{i}{\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{e}^{-t^2}}{x - t + iy}dt \tag{62}$$

Multiple efficient computational approximations have been developed for this function and are used in practice [68]–[70]. In the hapi [27] code the approach proposed by Schreier et al. [70] has been adopted. As such, the computation of the Voigt profile is typically performed as follows:

$$I_V\left(x,y\right) = \frac{1}{\pi}\mathrm{Re}\left[w\left(x,y\right)\right] \tag{63}$$

The difference between the Voigt and Lorentz profiles is visualized in Figure 17. The graph was computed using hapi [27] for $CO_2$ at 50°C at 1 atm pressure. The region of the changes is best visible in the lower part of the plot, where the difference in the profiles is depicted. The absolute peak height and the slope of the peak are most affected. The difference between the two line shapes increases with decreasing pressure.

While the Voigt model is standard for many applications, its shortcomings have been thoroughly presented in [31, p. 79ff]. Since the measurement resolution has dramatically improved over the last decades, more accurate modeling is needed. While many different models were created to incorporate the effects of velocity-changing collisions and their speed dependence and speed dependence of their shifts, it took some time to define a standard model recognized in the community. This was an especially crucial step since the line databases only provided Voigt parameters for a long time. The higher precision model that finally emerged is the so-called Hartmann–Tran (HT) model or Partially–Correlated Quadratic–Speed–Dependent Hard–Collision (pCqSDHC) profile [71], which will be described in the following section.

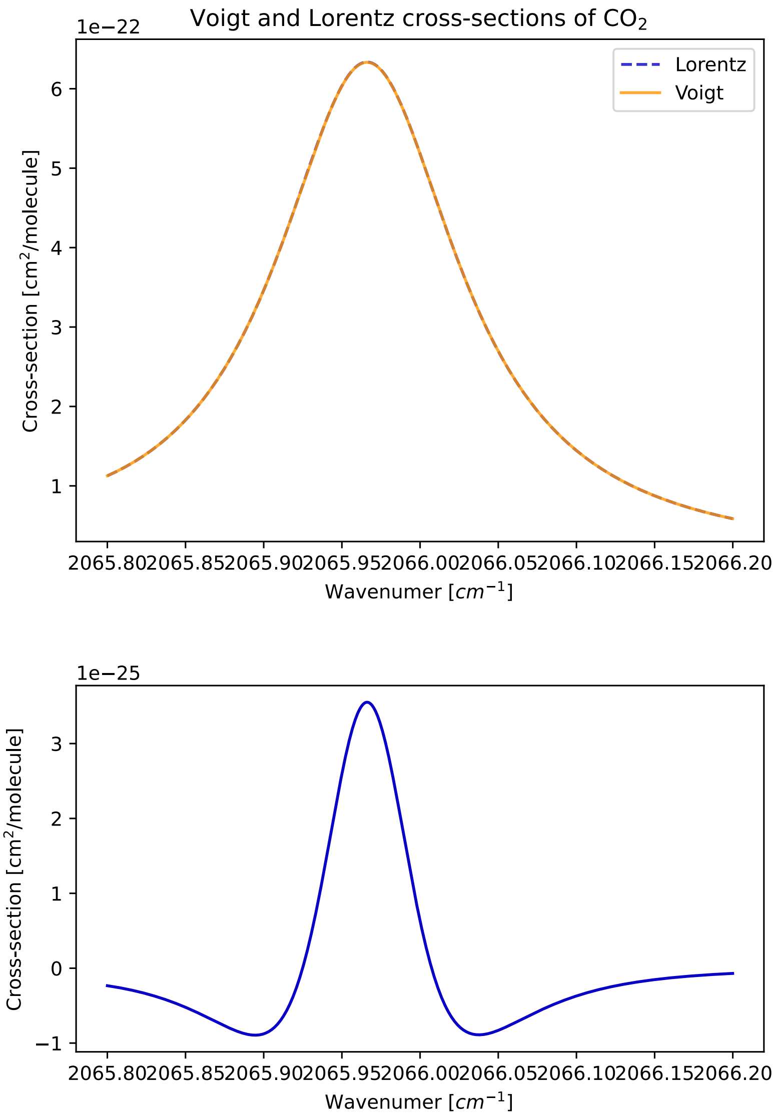


Figure 17: Difference of the Voigt and Lorentz profile of $CO_2$ at 50°C and 1 atm. The top plot shows the two simulated peaks. Here the difference is almost to small to notice. Therefore, the lower plot shows the difference between the simulations, enhanced by three orders of magnitude.

### 4.1.2 Hartman-Tran Model

After the standardization process described in [71], the HT or pCqSDHC model has also been adapted by HITRAN [14] and hapi [27] which now provide the parameters needed to compute this Non–Voigt profile whenever possible. The advantages of this line shape model have been summarized in [72] as:

- sufficiently physically based
- meaningful line parameters
- just slightly larger compute time
- many other profiles as limit cases
- compatible with line mixing effects

Having multiple other profiles as limit cases allows the usage of the model even if some required parameters remain unknown as the model then reverts to other profiles. Those are further described later on. The model includes multiple physical effects in addition to Doppler and Lorentz broadening by dephasing collisions. The pCqSDHC model considers collision-induced velocity changes, speed-dependent broadening, shifting coefficients, and a correlation between velocity and internal-state changes [72]. It thereby expands the previously described Dicke narrowing 2.3.4 by considering the speed dependence of this effect. To fully integrate all those effects, seven parameters are needed. Those are presented in Table 1.

| Parameter | Name | Unit |
|---|---|---|
| $\nu_0$ | Unperturbed position of the transition | $[cm^{-1}]$ |
| $\gamma_D$ | HWHM of Doppler broadening | $[cm^{-1}]$ |
| $\gamma_0$ | HWHM of collisional broadening (speed–averaged) | $[cm^{-1}]$ |
| $\gamma_2$ | Speed dependence of the line–width | $[cm^{-1}]$ |
| $\Delta_0$ | Speed–averaged line–shift | $[cm^{-1}]$ |
| $\Delta_2$ | Speed dependence of the line–shift | $[cm^{-1}]$ |
| $\nu_{VC}$ | Velocity–changing frequency | $[cm^{-1}]$ |
| $\eta$ | Correlation parameter | [1] |

Table 1: Line shape parameters for the computation of the HT-profile

| Profile | Reference | Limit of the PCqSDHC profile |
|---|---|---|
| Voigt | [26] | $\nu_{VC} = \eta = \gamma_2 = \Delta_2 = 0$ |
| Rautian | [40][ | $\eta = \gamma_2 = \Delta_2 = 0$ |
| quadratic speed dependent Voigt | [73] | $\nu_{VC} = \eta = 0$ |
| quadratic speed dependent Rautian | [40] | $\eta = 0$ |

Table 2: Limit cases of the PCqSDHC profile.

The profile is computed as [27, Version 1.1.0.9.7]:

$$I_{pCqSDHC}(\nu) = \frac{1}{\pi}\mathrm{Re}\left[\frac{A(\nu)}{1 - [\nu_{VC} - \eta(C_0 - 1.5C_2)]\, A(\nu) + (\eta C_2)B(\nu)}\right] \tag{64}$$

with $A(\nu)$ and $B(\nu)$ defined as:

$$A(\nu) = \sqrt{\pi}\frac{\sqrt{\ln(2)}}{\gamma_D}\left(w\left(iZ_1(\nu)\right) - w\left(iZ_2(\nu)\right)\right) \tag{65}$$

$$B(\nu) = \frac{1}{\tilde{C}_2}\left[-1 + \frac{\sqrt{\pi}}{2\sqrt{Y_c}}\left(1 - Z_1^2\right)w\left(iZ_1\right) - \frac{\sqrt{\pi}}{2\sqrt{Y_c}}\left(1 - Z_2^2\right)w\left(iZ_2\right)\right] \tag{66}$$

where $w(x)$ is the complex probability or Voigt function defined as [67, p. 297ff]:

$$w(z) = \frac{i}{\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{e}^{-t^2}dt}{z - t} \tag{67}$$

The other variables are defined as:

$$Z_1(\nu) = \sqrt{X + Y} - \sqrt{Y_c} \quad \text{and} \quad Z_2(\nu) = \sqrt{X + Y} + \sqrt{Y_c} \tag{68}$$

$$X(\nu) = \frac{1}{\tilde{C}_2}\left[i\left(\nu - \nu_0\right) + \tilde{C}_0\right] \tag{69}$$

$$Y(\nu) = \frac{1}{\left(2\frac{\sqrt{\ln(2)}}{\gamma_D}\tilde{C}_2\right)^2} \tag{70}$$

with:

$$C_0 = \gamma_0 + i\Delta_0 \quad \text{and} \quad C_2 = \gamma_2 + i\Delta_2 \tag{71}$$

$$\tilde{C}_0 = \nu_{VC} + (1 - \eta)(C_0 - 1.5C_2) \quad \text{and} \quad \tilde{C}_2 = (1 - \eta)C_2 \tag{72}$$

Using a combination of Voigt functions $w(z)$ dramatically reduces the computational power needed for this model as the same approximations described in Section 4.1.1 can be used. The limit cases of the pCqSDHC model are described with corresponding references in Table 2. Those models can, therefore, easily be modeled by setting the corresponding parameters to zero. As such, a meaningful line profile is still provided for many parameter settings with unknown parameters.

## 4.2 Line Databases

The parameters used in absorption modeling are collected in so-called line databases. They are typically curated and include or adapt parameters only after a quality assessment. Most databases can be found in the virtual atomic and molecular data center [74], [75], which aims to unify access to this kind of data. Several line list databases are available there. They are specialized for different applications and use cases. Bernath [76] comprehensively compares the various databases available and their application cases. It is essential to keep in mind the different units used in different databases and their conversion as well, as no standardization between the databases has been reached. The HITRAN database [14] was chosen for all subsequent experiments and simulations as it covers a wide range of spectral regions and molecules and broadening parameters. For those reasons, it is the most common database in trace gas sensor development. Parameters in HITRAN are based on a combination of precise spectral measurements and physical chemistry. It is frequently updated and extended. The database currently holds line data for 55 molecules. Each parameter also contains an uncertainty associated with it, which can be relatively high. Typically, broadening parameters for self and air are contained. Only a few molecules include additional broadening parameters (for example, water has $CO_2$ broadening parameters). As such, the full HT model can hardly ever be used with data available from HITRAN. It also has to be remembered that the simulation parameters are not necessarily correct. Even though great care is taken by the curators, some parameters are still somewhat vague and thus can cause misinterpretations if too much weight is given to the simulated data.

## 4.3 Simulation Code

In combination with the HITRAN online web interface [77], the simulation and interface code in Python called hapi [27] was introduced. This implementation has been continuously improved and includes a download Application Programming Interface (API) and the spectral modeling functions presented previously. With the 2020 update to the HITRAN database [14], a new version of hapi, hapi2, was introduced, which is intended as a complete do-over of the API. Currently, it covers only the download API but is intended to include the full hapi functionality. The work in this thesis has been performed with the original hapi code (Version 1.1.0.9.7). Still, a transition to hapi2 is advisable for future work once the full functionality is implemented, as the code is expected to be cleaner and faster due to the improvements.

HITRAN and hapi often use the so-called HITRAN parameters, which refers to the presentation of the absorption cross-section in $\mathrm{cm}^2/\mathrm{molecule}$, while other databases and codes typically provide $\mathrm{cm}^{-1}$. A conversion can be performed by multiplying the HITRAN units with the volume number density $\rho$ in $\frac{\mathrm{molecules}}{\mathrm{m}^3}$. The conversion is provided in Equation 73 and depends on the current environmental settings of pressure $p$ and temperature $T$. The Boltzmann constant $k$ is needed for conversion as well. Notably, inside hapi, the Centimeter-Grams-Seconds unit system is used.

$$\left[cm^{-1}\right] = \left[\frac{cm^2}{molecule}\right] \rho = \frac{p}{k_B T} = 7.3389399 e^{21} \frac{p}{T} \tag{73}$$

## 4.4 Limitations

Even though the line based simulations have greatly improved in the last decades, the following limitations have to be considered when using them:

- Not all molecules are available
- Not all transitions are available
- Parameters are not without error
- Not all parameters are available for each transition
- Broadening can mostly only be simulated for synthetic air and self
- Depending on the parameters and algorithm used, not all physical effects are considered
- Parameters are often provided with a big error margin
- Extreme environmental circumstances necessitate different simulation approaches

While some of the mentioned limitations are easily accessible during experimentation, such as the unavailability of specific molecules, others are often overlooked by practitioners. The broadening effects of air, for example, are often used as a proxy even though the actual broadener is vastly different. This is, for example, the case for broadening in a natural gas deposit where $CH_4$ and other hydrocarbons are most prominent. Therefore, careful and mindful use of the available tools is required. Due to those limitations, additional databases containing cross-sections of well-known molecules and mixtures are needed. Those are presented in the following section.

## 4.5 Cross-Section Databases

Precise measurements are often employed for molecules where simulation is not amendable, but that are of great interest in research. The most extensive compilation of those measurements, converted to absorption cross-sections in $\frac{\mathrm{cm}^2}{\mathrm{molecule}}$ is available from HITRAN [78] including also a large portion of the Pacific Northwest National Laboratory (PNNL) library [28]. The measurements are available through HITRAN online [77] and include the original reference data. Currently, more than 2000 spectra from 328 molecules are available. Typically, multiple measurements at different environmental settings were collected at comparatively high resolution (around 0.5-0.1 $\mathrm{cm}^{-1}$). Those can be incorporated with simulated data to obtain approximate mixture spectra.

Care must be taken to ensure correct data handling when using measured absorption cross-sections in simulation. Measured cross-sections typically contain measurement artifacts that cannot be prevented. Those are commonly noted in the original publication, for example, in [79] and often include fringes from the measurement device and spectral residuals from other molecules, mainly water. Those need to be kept in mind when using those measured absorption cross-sections in downstream tasks, as especially machine learning approaches could overfit on those artifacts. In addition, the noise in regions without absorption should not be underestimated. Therefore, those measured cross-sections are not easily scaled to high concentrations. The maximum sensitivity on each absorption measurement has to be considered separately for every original publication, but cross-sections of less than $10^{-25}\,\frac{\mathrm{cm}^2}{\mathrm{molecule}}$ can typically not be considered applicable. Some original publications have also published the measurements from which the final absorption cross-section was derived. Those measurements can be used to include additional variation in the data. Finally, measured absorption cross-sections are only available at specific configurations of pressure and temperature, and interpolation is not readily available.

## 4.6 Other Approaches

In addition to the presented databases for spectral data, simulation from line data, and measured absorption cross-sections, additional techniques exist. For the simulation of absorption cross-sections, ab initio approaches, which do not include any parameters like collisional broadening, are also available. Those simulation approaches have not yet reached a quality to be of interest for their use in sensor development or industry, but an overview can be found in [31]. Another interesting approach has been followed by Michalenko et al., who used a neural network to predict some of the line parameters for atomic spectra [80].

Finally, Toon et al. [7], [15] have developed Estimated-Pseudo Line Lists (EPLLs) at the Jet Propulsion Laboratory (JPL) to allow environmental adaption of measured absorption cross-sections to different pressure or temperature configurations. To create an EPLL, the partition function of the molecule in question needs to be known, and multiple high-precision measurements of the molecule in question at different environmental configurations are required. With this knowledge, an over-constraint physical model is created. This model corresponds to a devised line list with evenly spaced lines on a wavenumber grid. For each pseudo line, the lower state energy E” and its strength are fitted from the measurement spectra employing the Voigt line shape function. Standard simulation can now be conceived from this pseudo line list, leading to absorption cross-sections at new environmental configurations. An extensive collection of EPLLs can be found at [81] including documentation on their generation process. EPLLs have been successfully employed in atmospheric composition research [82], [83]. Their advantages include compensation for measurement impurities, as multiple measurements are used for the generation and easy handling in combination with line-data bases. Their disadvantage includes the need for a large amount of high-precision measurements and precise knowledge of the molecule, including the partition function.

The presented simulation approaches and databases provide the underlying data for the results in part III. In Chapter 6, a deep learning approach to achieve cross-section adaptation to different pressure configurations is presented, which provides an alternative to EPLLs and uses hapi simulated absorption spectra as the training data. Similarly, the synthetic photoacoustic spectra described in Chapter 7 use simulated or measured absorption cross-sections as input parameters.

# 5 Machine Learning and Chemometrics

In this chapter, the machine learning techniques applied in the result section, Chapter 7 and the deep learning models used in Chapter 6 are introduced. An overview of the field of chemometrics and the associated machine learning methods with a focus on absorption spectroscopy is provided. After a short introduction of chemometrics and Artificial Intelligence (AI) in general, where we classify the techniques covered in this thesis, a more in-depth overview of supervised methods in gas spectroscopy will be given. Afterward, essential practices in machine learning focused on spectral data are described, including preprocessing, data augmentation, and hyperparameter tuning. Next, an overview of variable selection techniques focused on spectral data is provided. Finally, the adaption of spectral data with the help of machine learning is presented. Here, the different spectral metrics available to compare the similarity of spectra are highlighted.

While chemometrics is a broad topic, including the interpretation of all chemical signals, here we only focus on the interpretation of spectra, namely absorption spectra. The field is partially visualized in Figure 18, indicating its broadness. Chemometrics can be defined as the use of mathematical and statistical methods to analyze chemical data and design measurement setups [84]. Chemometrics includes the analysis of spectral data and covers all kinds of chemical applications, like molecular synthesis and experiment design. When handling spectral data, a broad range of measurement techniques and data types remain. So not only absorption spectra but also Nuclear Magnetic Resonance (NMR) or Raman spectra can be handled by chemometric methods [85], [86]. Even though machine learning techniques applied to those spectroscopic methods can be similar and used as inspiration, the underlying physical principle of the data differs for each kind of spectroscopy. Therefore, a direct transfer of machine learning models or even approaches is not always possible [87]. Even at the most fundamental level, absorption spectra still show substantial differences when taken in different physical states. For example, spectra acquired from solid or fluid materials suffer from strong scattering effects, which is not the case for gaseous samples [88]. Spectral data from different measurement techniques or materials might appear very similar to the untrained observer but require a very different treatment during analysis.

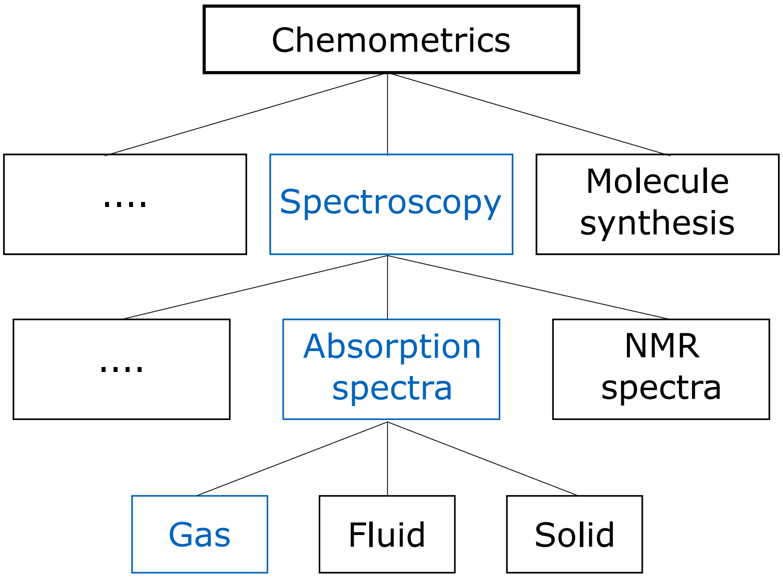


Figure 18: Schematic overview of the field of chemometrics and the allocation of gas absorption spectroscopy.

While chemometrics started at the forefront of pattern recognition and, thereby, machine learning in the 1970s, Brereton, one of its most recognized current leaders, describes it as a field that strongly diverged from the mainstream machine learning and pattern recognition community, focusing on evolving their own, highly specified tools with little communication to other disciplines [89]. The field has been extensively reviewed in a series of publications. From its beginnings in the 1970s [90], [91], followed by the review from the period from 1985-1995 focused on multivariate analysis [92]–[94] and more recent advances in supervised multivariate analysis [95], [96]. Due to the divergence of the fields, early adapters of typical machine learning technologies in absorption spectroscopy have often provided tutorials focused on a translation between the two disciplines [87], [97], [98]. Even though larger deep learning models like Convolutional Neural Networks (CNNs) and Transformer models have been applied in absorption spectroscopy, the amount of data available for training remains often at a few thousand samples, which results in a strong risk of overfitting and limits the application of large models [99], [100]. The risk of overfitting will be further described in the context of neural networks in Section 5.3.

## 5.1 Artificial Intelligence and Machine Learning

The influence of Machine Learning (ML) and AI on the field of chemometrics has grown in recent years. This increase is visualized in Figure 19. The topic of ML in chemometrics increased from around 12.5 % in 2014 to 30 % in 2022. Data on the number of publications was acquired from apps.dimension.ai using a free-text search for the keywords "chemometrics", and "chemometrics" and "machine learning" simultaneously. Since AI and ML cover an extensive range of techniques and applications, their full description is beyond the scope of this thesis. Here, only the techniques used within the thesis will be described, and a focused overview of the applications in absorption spectroscopy will be provided. For a complete overview of the field and a fundamental understanding of the techniques, the inclined reader is referred to excellent books on the topic [101]–[103].

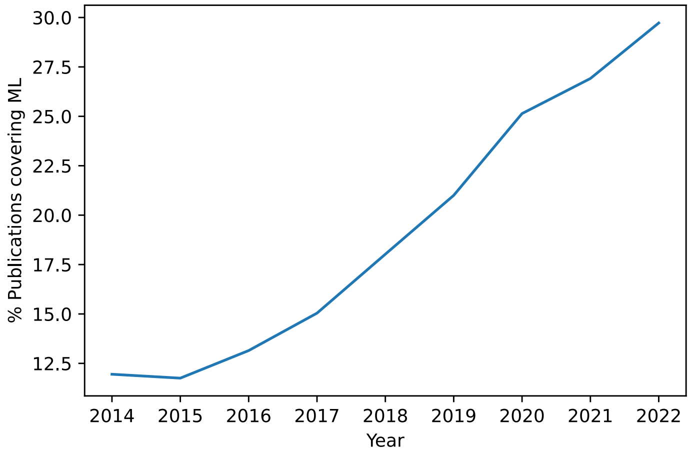


Figure 19: Percentage of publications containing the keyword "Chemometrics" and also "Machine Learning". (created using apps.dimension.ai)

As AI and ML are umbrella terms, each covering a broad field, a more precise definition for this thesis is required. The research community has not yet conceded to one full definition [104]. In this thesis, we follow the commonly used definition of the Oxford dictionary [1] and treat AI as a vast field that is defined by algorithms that perform tasks that usually require human intelligence. ML itself is a subset of AI, as it is considered to achieve this intelligence by extracting knowledge from data without being explicitly programmed. Within ML, three main types of learning are commonly differentiated, which are visualized in Figure 20 [103, Chapter 18].

Supervised learning is the most common type of learning, referring to a situation where the input and associated output data are available. This refers to, for example, having a fully labeled spectroscopic dataset where the actual gas composition is known for each acquired spectrum. On the other hand, unsupervised learning refers to the training situation where this labeling information is unavailable. In spectroscopy, this can, for example, be the case in quality monitoring, where the actual chemical composition of the analyte is not known, and only deviations from the norm need to be analyzed. For this example task and many other cases, those two types of learning can overlap. It can often occur when only a few labeled samples are available, but a vast amount of unlabeled data can be acquired at a lower cost. This is then considered semi-supervised learning. An extensive range of methods to combine the two kinds of learning exists. Finally, reinforcement learning is applied to an entirely different situation. Here, an environment and an agent interact to achieve a good acting policy, which can maximize the agent's reward. This is commonly known as the setup of gaming scenarios, for example, in the Chinese game of Go (with the agent Alpha Go) and other games. In spectroscopy, this kind of learning is less common,

[1] https://www.oxfordlearnersdictionaries.com

but techniques from reinforcement learning might be used, for example, in experiment design. As this thesis focuses on the application of supervised learning in gas absorption spectroscopy, an overview of applications and approaches will be provided in the next chapter.

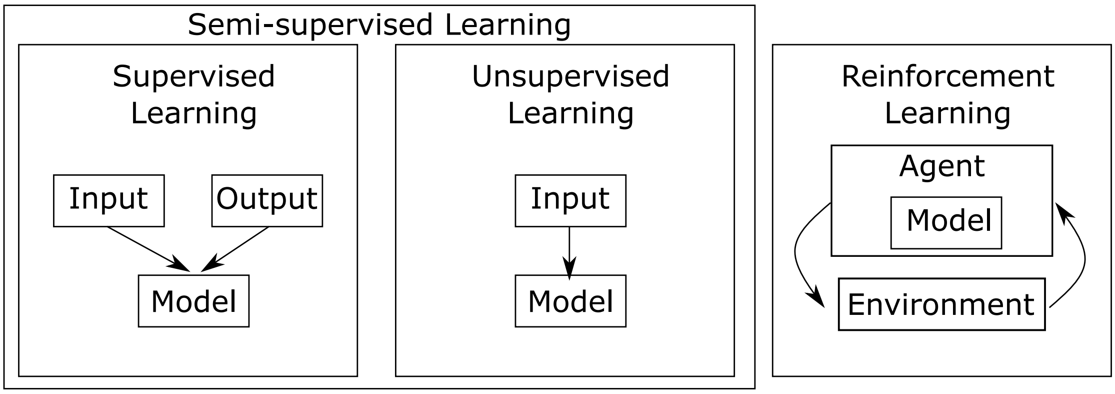


Figure 20: Overview of the different types of machine learning.

## 5.2 Supervised Learning - Algorithms

The field of supervised machine learning and the available range of algorithms is very vast and has grown tremendously in the last few years. Many approaches from ML were applied in spectroscopy. Therefore, the focus will be on models commonly used and applied within this work. The selected algorithms are summarized in Table 3, including their essential characteristics like linearity, ease of interpretation, and if they are considered an ensemble method. After introducing the respective models, examples of their application in (gas) spectroscopy will be presented in the following sections. Due to the vast number of models, only a broad introduction to each model is provided. Citations for a more in-depth discussion and further explanations of the models are provided in each respective section. Due to their importance and diversity, neural networks will be covered in a separate section.

| Model | Non-linear | Ensemble | Interpretation | References |
|---|---|---|---|---|
| MLR | – – | – | + | [101, Chapter 3.1] |
| MCR | – – | – | ++ | [105], [106] |
| PLS | – | – | + | [107], [108] |
| SVR | + | – | – | [101, Chapter 7.1.4] |
| RF | ++ | + | – | [101, Chapter 14.4] |
| GB | ++ | + | – | [109, Chapter 10] |
| ANN | ++ | – | – – | [102] |

Table 3: Overview of machine learning approaches described and their characteristics.

**MLR** Multi Linear Regression (MLR) is the most basic linear model, which can be seen as an extension of simple linear regression. As presented in Equation 74, it is a linear combination of the adapted weights $W$ and the input variable $x$. The weights are typically fitted using, for example, least-squares regression. An additional component can extend the input vector $x$ to provide a bias term that is not dependent on the input variable. An in-depth description can be found in [101, Chapter 3.1]. MLR does not cope well with the high collinearity of spectral data, leading to poor performances in those cases.

$$f(W,x) = W^T x \tag{74}$$

**MCR:** Multi Curve Regression (MCR) has been specifically designed to tackle the mixture problem, where multiple spectral signals overlap. It does so by recreating the measured spectrum through a linear weighted combination of single-component spectra weighted by their estimated concentration. Errors of this fitting are indicated by the error term, which is minimized. This represents a bilinear combination of already-known base spectra. Many fitting algorithms can execute the optimization, but least-squares remains the most popular approach. An in-depth tutorial and review on techniques and best practices by de Juan et al. can be found here [105] and here [106]. This method incorporates physical information in the form of base spectra, thus providing a straightforward interpretation. This makes it highly dependent on the quality of those spectra and does not easily allow for an application when additional, changing, or unknown components are included.

**PLS:** Partial Least Squares (PLS) or Partial Least Squares Regression (PLSR) can be considered an extension of MLR, which allows for the use of latent variables and adds the optimization of the correlation of those latent variables. It thereby can cope better with the high collinearity of spectral data. PLS can be considered the workhorse of chemometric analysis [107]. The PLSR is linear in its parameters. A non-linearity can be included via a non-linear kernel function, which transforms the input space. The usage of kernel functions is not very common in chemometrics. The difference to MLR stems from the different optimization of the weights. In the first step, the correlation of the latent variables with the target variables is optimized. Next, the loading vector is estimated. The number of latent variables in the model needs to be optimized during hyperparameter tuning. Multiple in-depth tutorials focused on spectral data exist [107], [108]. PLS results can be interpreted by analyzing the loading and score vectors.

**Support Vector Regression:** While Support Vector Machine (SVM) for classification is based on maximizing the margin between the classes, Support Vector Regression (SVR) has a small margin around the correct prediction, which is not penalized. Similar to SVM, SVR is typically extended with a kernel function, which transforms the input parameters to a different space, which might allow for a better regression. An in-depth explanation can be found in [101, Chapter 7.1.4]. Using a kernel function allows the SVR model to go beyond the linear modeling and integrate non-linear behavior. The kernel function needs to be tuned during hyperparameter tuning.

**Random Forest:** Random Forest (RF) is an ensemble model which combines multiple Decision Trees (DTs). A DT is a simple, sequential model that divides the input space. It can be adapted for use in regression. This results in a non-linear, step-wise estimation. The model and fitting process are described in depth in [101, Chapter 14.4] and [103, Chapter 18.3]. While those tree models are easily created, they are not very expressive. Therefore, an ensemble of DTs, called RF, is used. Here, many DTs are fitted on different subsets of the training data and input attributes. Their predictions are averaged to generate the final output. With the ensemble, outliers in the data can be compensated for. Hyperparameters for pruning the underlying DTs and the combination of DTs into a RF need to be tuned. While RFs bring all the positive aspects of an ensemble method, the underlying DT models suffer from problems with interpolation between training data points as they were originally intended for classification.

**Gradient Boosting:** Gradient Boosting (GB) is rather similar to RFs as it is an ensemble method built on DTs as the base model. It differs from RF in the generation of the model. While in RF, each tree is constructed independently, and the results are aggregated, in GB, each DT depends on the result of the previous iteration, and the results are combined one at a time. An in-depth explanation of GB is provided in [109, Chapter 10]. The hyperparameters of the pruning of the DTs and the GB generation need to be tuned. The advantages and disadvantages are similar to those of RFs.

## 5.3 Neural Networks

Artificial Neural Networks (ANNs) form the basis of many new applications around ML and have dominated the developments in the last decade. Therefore, a more in-depth discussion of the different architectures and techniques surrounding neural networks is provided in this thesis.

Even though ANNs are inspired by the architecture of brain neurons, they significantly differ and fail to incorporate essential features of our brain. They are now ubiquitous in many domains, including chemometrics. The simplest kind of neural network is a concatenation of weighted functions, including an activation function. Those neurons can be combined in layers, and multiple layers form a network. The weights within those networks are adapted to the training data using backpropagation. Any function can theoretically be fitted with enough neurons, which renders the ANNs universal function approximators. The more adaptable parameters or weights are available to the network, the more samples must be contained in the training data to counter overfitting. Overfitting specifies a too-perfect network performance on the training dataset, failing to generalize well on the validation and test dataset. It can be colloquially compared to having learned the training samples by heart and not being able to apply the pattern behind the data. Many techniques and specialized layers have been developed to improve neural network performance and counter overfitting. An in-depth overview of neural networks and essential techniques can be found in [102].

Like other ML techniques, multiple training parameters of ANNs need to be tuned. Various methods exist, which are only partially described in standard references [102], as they are continuously developed and improved upon. In addition to the training process, the architecture of the network can be tuned. Here, a wide variety of possible layer types can be used. We only cover the layers applied within this work, as a full review would be out of scope:

- **Fully Connected**: A standard fully connected layer consists of weights $W$ and an activation function $\Phi$ as presented in Equation 75. An additional component can extend the input vector $x$ to provide a bias term that is not dependent on the input variable. Each neuron of the incoming layer is connected to each neuron of the outgoing layer via a weight that can be adapted during training. The choice of activation function depends on the task at hand and serves to integrate the non-linearity in the network. Often, a Leaky Rectified Linear Unit (ReLu) function is applied.

$$y = \Phi(Wx) \tag{75}$$

- **Convolutional Layer**: A CNN contains convolutional layers which became popular in image recognition [110]. Those convolutional layers perform a discrete convolution on a window from the input data as presented in Equation 76 and slide over the full range of input data. This way, the convolutional filter weights are shared for the full spectrum. While most examples of convolutions are used in image processing and apply to two-dimensional data, their application to one-dimensional spectral data is not uncommon. The learned convolutional filters have been found to resemble typical spectral preprocessing steps like differentiation [97]. It is often combined with batch normalization or pooling layers described in [102].

$$y = \frac{1}{x}\sum_i w_i x_i \tag{76}$$

- **Residual Blocks**: When the model depth is increased beyond a dozen layers, the vanishing gradient problem can cause the training to fail. To counter this continuous shrinking of the backpropagated gradient through multiple layers, the residual connection was introduced by He et al., allowing much deeper architectures [111]. A residual connection can be considered a direct routing where the input is added to the output. This creates a residual block, which can include multiple different layers. A schematic of this block is visualized in Figure 21.

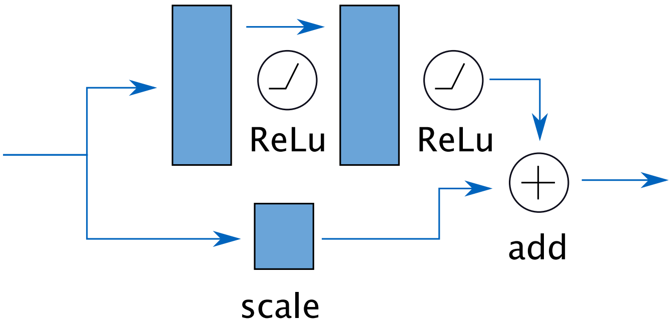


Figure 21: The basic schematic of a Residual block. The input is split into two paths, one passing through two layers, commonly convolutional layers, while the other is scaled and added to the output of the first path before leaving the residual block.

- **Squeeze and Excite Blocks**: Squeeze and Excite (SE) connections as presented in [112] have been used within this work. They have been inspired by recent transformer networks adding cross-channel connections to the architecture. It is visualized in Figure 22. A global average pooling operation extracts the channel mean. Those averages can interact in a very small two-layered linear fully connected neural network before being reintroduced to the full block input via multiplication. This way, a scaling of each channel dependent on the other channels is achieved.

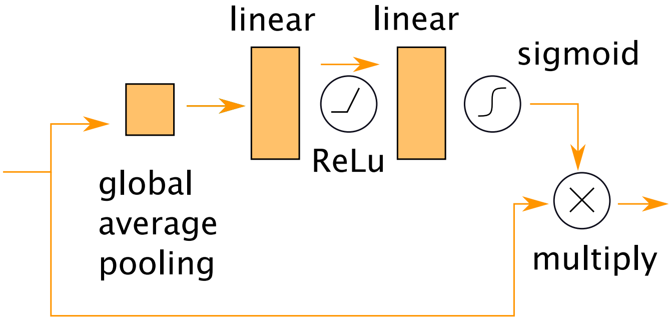


Figure 22: The basic schematic of a Squeeze and Excite block. The input is split into two paths. The upper one passes through a global average pooling to extract the channel means, which are fed to a two-layer network. The output is scaled by a sigmoid function and multiplied by the input data from the second path.

Two more important concepts specific to neural networks need to be presented for the remainder of this thesis. The first is the **Latent Representation**. Layers within the network are referred to as latent layers as they contain a transformed version of the input data after activation, which is not easily interpretable. This latent information can nevertheless be used for specific applications. This includes regression on the latent representation, denoising, anomaly detection, and compression [113]. It is typically obtained using an Autoencoder (AE) architecture as depicted in Figure 23. This unsupervised architecture is hourglass-shaped with respect to the parameters, with a bottleneck in the middle. The same data point fed into the network is used as the target, so the AE learns to reconstruct the data. The latent information in the bottleneck can then be extracted and used [113]. There are many expansions around the basic AE, including Variational Autoencoders (VAEs) or Wasserstein-Autoencoders, as well as combinations of Generative Adversarial Networks (GANs) and AEs. For a more in-depth discussion, the inclined reader is referred to [113]

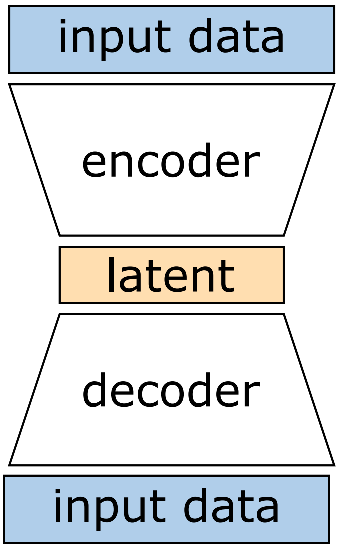


Figure 23: The basic architecture of an autoencoder used to create an expressive latent representation. The output data of an autoencoder corresponds to the input data. The encoder compresses the data to fit the latent representation. The decoder then reconstructs the input data from the latent representation.

Finally, the concept of **Fine Tuning** is one of the most successful approaches to counter overfitting in the case of small specialized datasets. It refers to adapting a previously trained network on the final, specialized dataset. Therefore, the network is initialized with the weights of the previous training and trained for a few additional epochs on the small, specialized dataset. This fine-tuning can be applied to the full network or only to the last few layers, depending on the domain gap between the previous and final datasets. This allows using the knowledge the model has previously acquired, especially in the first few layers, which are often quite general, to be used in the final task. This approach is common in chemometrics and often applied in the context of calibration transfer [114], [115].

## 5.4 Supervised Learning - Applications

ML in absorption spectroscopy already spans a vast field. Therefore, here we focus on an extensive review of the application of machine learning techniques in photoacoustic multi-component gas spectroscopy while highlighting important works and reviews that cover the broader area. An overview of essential publications applying machine learning to PAS is summarized in Table 4, which will be described in more detail in this section.

The first multi-component PAS detection systems were devised for freon-125 and acetone quantification by Lewicki et al. in 2007 [116] and further developed by Kosterev et al. in 2010 [117]. While those two publications show photoacoustic spectral measurements, which demonstrate the separability of the spectral components via MCR, they lack a rigorous evaluation of the proposed concept over a broad range of concentrations. The concentration error was not quantified during either experiment. They mainly present the feasibility of the MCR approach on Amplitude Modulation (am) QEPAS measurements over a relatively broad spectral range (1196-1281 cm$^{-1}$) with no non-spectral interferences.

| **Reference** | **Algorithm** | **Approach** | **PAS Technique** | **Analytes** |
|---|---|---|---|---|
| [116] | MCR | Regression | AM QEPAS | Freon 125, Acetone |
| [117] | MCR | Regression | AM QEPAS | Freon 125, Acetone, Ethanol, water |
| [118] | PCA + SVM | Classification | LaserBreeze | LC and COP |
| [119] | SVM | Classification | LaserBreeze | four illnesses |
| [120] | PCA + SVM | Classification | LaserBreeze | Myocardial infarction |
| [121] | MLR | Regression | WM QEPAS | Methane, Ethane |
| [122] | PLSR | Regression | WM QEPAS | Methane, Ethane |
| [123] | MCR | Regression | AM QEPAS | $CH_4, N_2O$ |
| [124] | PLSR | Regression | WM QEPAS | $CO_2, N_2O$ and $CH_4, C_2H_2, N_2O$ |
| [125] | MLR | Regression | AM PAS | Acetone, Ethanol, Water |

Table 4: Publications applying machine learning approaches in PAS.

Another interesting development was the creation of the photoacoustic spectrometer called LaserBreeze [126], which was developed and used for exhaled breath analysis. It was operated by a combination of two OPOs and could cover a range from 2.5 - 11 µm, applying am PAS. It was used in multiple studies that involved machine learning [118]–[120]. Those all combined a Principal Component Analysis (PCA) analysis and SVMs to classify illnesses. Those studies suffered from small datasets (72 samples [120], 61 samples [119], unknown [118]) and did not use an additional independent test set. Therefore, they did not provide actual insight into the performance of the proposed methods nor their applicability in the medical context.

Menduni et al. developed and evaluated different methane/ethane sensors using Wavelength Modulation (wm) PAS. A first sensor consisting of two light sources detecting peaks at 6047 and 5938 cm$^{-1}$ used a simple linear regression at the two peak positions to differentiate the two components [121]. A more complex interaction was presented in the second sensor analyzing peaks in the area from 2988-2991 cm$^{-1}$. Here, non-spectral interferences and the sensor saturation made quantifying the two components more complex and required a more intricate approach than MLR. For this reason, a PLSR was trained on 42 measurements to estimate the methane and ethane concentration. Special care was taken to make the system robust against changes in propane concentration, which also generates non-spectral effects [122].

Giglio et al. apply a MCR approach on previously measured spectra to separate the composite spectrum and quantify the concentration of $CH_4$ and $N_2O$ in the area from 1240-1320 cm$^{-1}$ [123]. They employed amplitude-modulated PAS and acknowledged that non-spectral interferences, for example, from an added water concentration, would require a more intricate quantification algorithm. Zifarelli et al. employed PLSR to differentiate and quantify mixtures of $CO_2$ and $N_2O$ on wm QEPAS in the spectral range from 2189 to 2191 cm$^{-1}$ and $CH_4, C_2H_2$ and $N_2O$ in the spectral range of 1295.5-1296.5 cm$-1$. Notably, they did not only train the PLSR on measured test sets but also employed data augmentation by training the algorithm on a linear combination of their measurements [124].

The works presented within this thesis, which were partially published, present an am PAS system in the wavenumber region from 1209-1210 cm$^{-1}$ for acetone, ethanol, and water quantification [125]. This work and an extension to other ML techniques will be presented in Section 7.1.2 and offer an in-depth comparison of different machine learning approaches and their interaction with synthetic or measured training data.

From the presented publications, it becomes evident that PAS spectral analyses applying machine learning are not the most common and currently only apply basic machine learning approaches. PAS is mainly used as a single-point measurement technique, focusing on one analyte in a region with little to no interference. Spectral measurements can provide additional insight but are more complicated to understand than simple absorption spectra, as non-spectral interferences can influence the signal additionally. When those problems can be overcome, as suggested within this thesis by using simulation and machine learning, PAS spectral measurements can be used to combine the advantages of PAS as a direct signal and high signal amplitudes thanks to lock-in-amplification, with multi-component analysis.

The photoacoustic non-spectral interferences are why PAS needs to be considered differently than pure absorption measurements, as integrated with standard FTIR instruments. Those typically can be handled by MCR as no additional non-linear effects occur. In this application, the advantages in ML techniques are mostly faster processing, especially when handling many components. Goldschmidt et al. present an ANN for the prediction of CO and $N_2O$ using a dual-comb spectrometer in the wavenumber range from 2186-2119 cm$^{-1}$. Notably, they trained their network on synthetically generated data, adapted for the deviations of the spectrometer in question, and reduced the inference time from 484s to 303µs [12]. On a similar note, Prischepa et al. [13] trained a neural network to detect multiple gas components but focused more on a criterion for the reliability of the prediction. Absorption spectroscopy on non-gaseous samples suffers from stronger interference and thus has much earlier started the adaption of machine learning methods. A review of the field can be found in [11], [127]–[129]. The algorithms presented in the previous section have been employed on non-gaseous samples in (photoacoustic) absorption spectroscopy, including SVR [119], [120], [130], ANNs [12], [131], and RF [99], as well as GB [132]. Also, more intricate neural network designs have been tested for the classification and regression of components. Those range from shallow CNNs [133]–[135] to deeper CNN architectures [99] and also include recent transformer models [100]. Applying large neural network models is often considered difficult, as only comparatively few sample spectra are used (few thousand) [11].

## 5.5 Common Techniques in the Field of Spectroscopy

As reproducibility in chemistry and chemometrics plays a crucial role, common methods must be established for machine learning experiments to ensure comparability. Due to the long separation of the fields [95], [96], a number of tutorials especially for chemists exist [87], [89], [97], [136]. This thesis will provide an overview of the commonly used methods along the model development chain. We will present dataset splitting, preprocessing, data augmentation, hyperparameter tuning, and evaluation of quantitative models focused on gas absorption spectroscopy.

### 5.5.1 Dataset Splitting

In chemometrics, small datasets containing many features are the norm. Therefore, additional care must be taken when splitting the dataset into training, validation, and test sets. Here, we follow the nomenclature of Bishop et al. [101, Chapter 1.3]. As the name implies, the training set is used for training the algorithm, while the validation set is used to adapt the hyperparameters of the training process or algorithm. Finally, the test set is only used after tuning to estimate the algorithm's performance on an unseen dataset. Typical splitting scenarios are visualized in Figure 24. Standard ratios for the splits are 70% training data, 20% validation data, and 10% test data. When very few samples are available, cross-validation can be used to increase the amount of training data available. The cross-validation setup is visualized in Figure 24 on the right. The number of cross-validation splits is a trade-off between training time and the available training data. The most extreme case is leave-one-out cross-validation, where only one sample is used as a validation split. Still, an independent test split is needed for the final evaluation. Another option is the so-called double cross-validation setup, separating the cross-validation into an outer and inner loop. This is visualized in the bottom plot c) of Figure 24. This way, each outer test split can be used for the final evaluation, as it was not used during hyperparameter tuning.

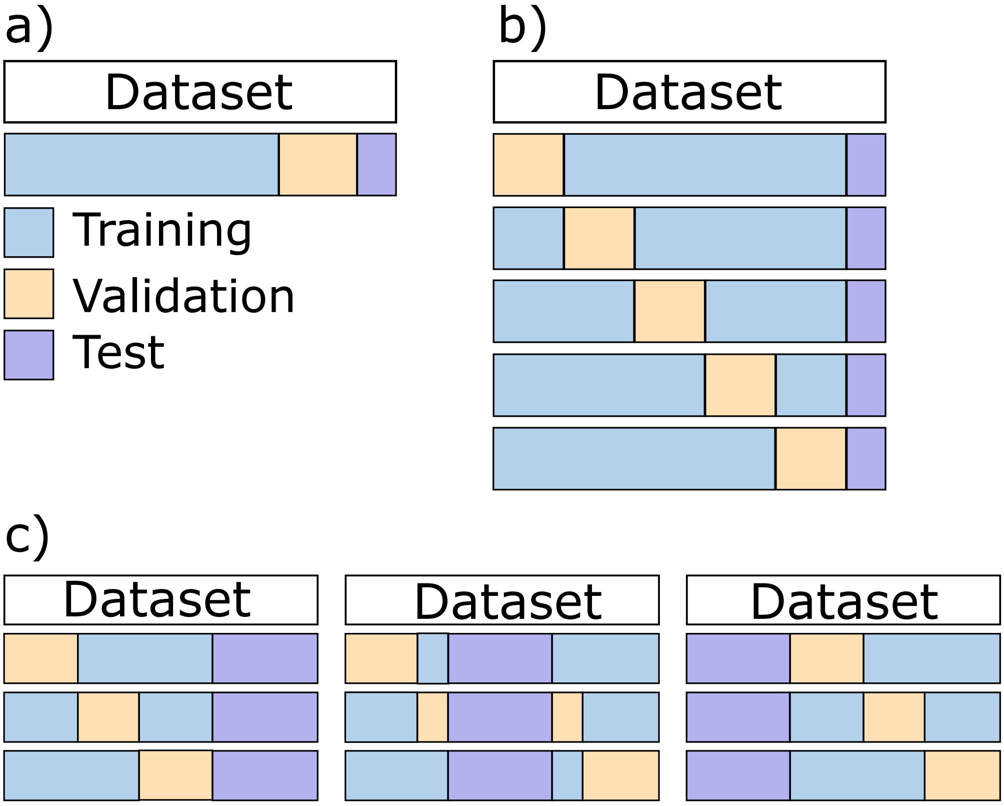


Figure 24: Visualization of dataset splitting. a) splitting into a train, validation, and test set, b) five-fold cross-validation with an independent test set, c) three-three double cross-validation.

Some attributes of the dataset have to be kept in mind when creating the splits:

- **Sample imbalance:** If the number of samples containing one analyte is much more frequent than another analyte, the splits must be stratified. In the training and validation splits, oversampling and weighted approaches can be applied to counter this problem.
- **Instrument variation:** If multiple instruments were used to create the dataset, those need to be either distributed equally over the splits or all of one instrument's measurements can be used as an evaluation set. The latter shows the possible transfer of the model to another instrument. Notable changes in the measurement setup, e.g., laser realignment, that significantly changes the available optical power, must be handled similarly.
- **Repeated measurements:** If the measurement of one mixture is repeated multiple times, those samples need to be kept in one split together. A repeated measurement in the evaluation split leads to increased scores that do not reflect the model's overfitting.
- **Preprocessing:** If data-based preprocessing is applied, only the training data can be used to adapt the preprocessing steps.

### 5.5.2 Preprocessing

Preprocessing corresponds to all data processing done before the training of the model. It is especially vital for absorption spectroscopy of liquids and fluids where scattering interferes with the signal. Many different preprocessing approaches exist and are regularly reviewed [137], [138]. We will exclude scatter correction from this section, as it is not applicable for gas spectroscopy. The inclined reader can find additional information on standard scatter correction methods like EMSC in the literature [138], [139].

Common preprocessing steps include:

- **Derivatives:** Spectral derivatives can be used to remove the baseline of a measurement. First or second-order derivatives are used depending on the background to be removed. A first-order derivative treats a constant background, while a linear background requires a second-order derivative. This preprocessing step can eliminate some information contained in the original data [138].
- **Savitzky-Golay derivative:** This moving window derivation fits an n-point polynomial to the data before the derivation. This counters outliers in the data while also providing a smoothing effect. The size of the moving window is a hyperparameter to be tuned [140].
- **Background subtraction:** Through a background measurement without the analytes, the background signal can be estimated and subtracted. If a background measurement is unavailable, the mean spectrum from the training dataset can be used as an approximation of the background [137]. In addition, it is possible to scale the data by the standard deviation to ensure values between zero and one [141]. Here, additional care is required as some existing scaling algorithms, for example, in sci-kit learn, perform mean scaling for the full data, not by wavelength.

- **Outlier removal:** Especially for shallow machine learning approaches, outliers can lead to grave errors. Therefore, outliers in a dataset can be removed before further processing. This is often done by finding outliers in the PCA space. One example can be found here [142].
- **Target scaling:** In multi-target analysis, it is crucial to have a similar numerical quantity to predict each component. So, the concentrations used as target variables must be transferred to a similar domain. This is often done via min-max scaling between zero and one.

Even though data preprocessing is less relevant for gases compared to scattering liquid and solid samples, it can lead to significant improvements. Nevertheless, selecting and combining the perfect preprocessing steps is tedious, and some combinatorial automatic approaches have been proposed [143]. When deep learning models, mainly CNNs, are applied, the effect of preprocessing is not as strong [11], [98]. Cui et al. have pointed towards the first layer of the CNN adapting to computations similar to common preprocessing steps [134].

### 5.5.3 Data Augmentation

Since large datasets are scarce and hard to obtain in spectroscopy, data augmentation yields great potential to enable the application of larger networks and other techniques from the field of deep learning [98]. Established methods to augment spectral data are described by Bjerrum et al. [141] and visualized in Figure 25. This includes adding a linear offset to the signal, multiplying it by a small factor, and changing the slope of the signal by a small amount. Finally, gaussian noise can also be added to the call.

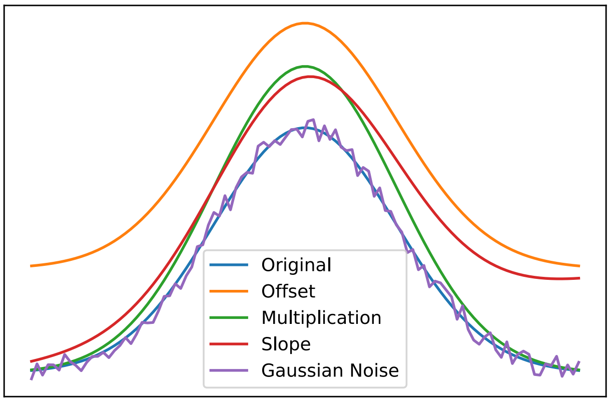


Figure 25: Different types of spectral data augmentation. Those can be applied separately or in combination.

Additional techniques are available for scatter-affected data, including liquids and solids, to augment it to different scattering variations. Recent approaches are based on physically grounded simulations [144].

To determine the strength of the augmentation methods, different techniques are available. The hyperparameter optimization can include those parameters, or the strength can be estimated from the training data [12], [141]. In the case of a background-free signal, a linear combination of the samples can also be used for data augmentation [124]. Each measured signal is scaled, and multiple signals are added to simulate different mixture concentrations. This approach can only be applied if the data follows an additive and linear trend, which is not the case for am PAS measurements affected by changes in the photoacoustic relaxation efficiency.

Finally, training data can be increased via simulation. This approach will be presented and discussed in this work for photoacoustic measurements and has also been used by Goldschmidt et al. in dual-comb absorption spectroscopy [12]. When using a physically grounded simulation, specific simulation parameters can be altered to provide a bigger variety of data, simulating, for example, a change in the background signal. This will be further explained in Section 7.

### 5.5.4 Tuning

A tuning of the algorithm's hyperparameters can significantly improve the performance of machine learning algorithms. This has been pointed out for spectroscopic data by Mishra et al. [11], and tutorials for spectroscopists to successfully implement hyperparameter tuning also exist [145]. Hyperparameters include not only the training settings of the machine learning model but can also include preprocessing steps and data augmentation parameters. A similar optimization of the model's architecture is called neural architecture search. Many tools implementing standard techniques exist. For example, in python, the libraries and services hyperopt[2] and Weights and Biases [146] are available. Simple tuning can also be performed directly in scikit-learn [147]. Bayesian methods are typically applied when many parameters must be tuned simultaneously.

### 5.5.5 Evaluation

A multitude of metrics exist to evaluate the performance of regression problems. Therefore, the metrics applied in this work and commonly used in literature will be described in this section. It has to be pointed out that in machine learning, the evaluation always takes place on the final test set unless stated otherwise. This highlights the need for a representative test set. In the following, the metrics are described, and their computation is provided in Equations 77 to 82. $y_i$ represents the algorithms prediction on the sample $i$ while $\tilde{y}_i$ represents the target and $\overline{\tilde{y}}$ is the mean of the target variable, $n$ is the number of samples.

- **MAE:** The Mean Absolute Error (MAE) presents the mean of the absolute deviation of the predictions to the targets. It is often preferred in chemical evaluations as it can be represented in the original units. Its computation is given in Equation 77.

$$MAE = \frac{1}{n}\sum_{i}^{n} |y_i - \tilde{y}_i| \tag{77}$$

[2] https://github.com/hyperopt/hyperopt

- **MSE:** The Mean Squared Error (MSE) presents the square of the deviation of the predictions to the targets. It is often used for training machine learning algorithms as it ensures purely positive errors while avoiding taking the absolute value. Its computation is given in Equation 78.

$$MSE = \frac{1}{n} \sum_{i}^{n} (y - \tilde{y})^2 \tag{78}$$

- **Huber:** The Huber loss is a combination of the MAE and MSE, which is intended to be more robust than any one of the two. It is linear for larger loss values (MAE) and quadratic for smaller loss values (MSE). The boundary $a$ can be chosen or adapted via tuning. Its computation is given in Equation 79.

$$Huber = \frac{1}{n} \sum_{i}^{n} \begin{cases} |y_i - \tilde{y}_i|, & \text{if } y_i - \tilde{y}_i > a \\ (y - \tilde{y})^2, & \text{otherwise} \end{cases} \tag{79}$$

- **RMSE:** The Root Mean Squared Error (RMSE) is an adaption from the MSE, taking the root of the MSE. This way, the interpretation of the squared signal is simplified. Its computation is given in Equation 80.

$$RMSE = \frac{1}{n} \sum_{i}^{n} \sqrt{(y - \tilde{y})^2} \tag{80}$$

- **MAPE:** The Mean Average Percentage Error (MAPE) sets the mean absolute deviation of the predictions to the targets in relation to the target's value. This metric is often applied when handling very different target values. Its computation is given in Equation 81.

$$MAPE = \frac{1}{n} \sum_{i}^{n} \frac{(|y - \tilde{y}|)}{\tilde{y}} \tag{81}$$

- R$^2$**-score:** The R$^2$ − score is the coefficient of determination. It quantifies the explained variance of the model. It is commonly used in spectroscopy, and its value ranges between zero and one, where one shows perfect performance. Its computation is given in Equation 82.

$$R^2 = 1 - \frac{\frac{1}{n} \sum_i^n (y - \tilde{y})^2}{\frac{1}{n} \sum_i^n \left(\tilde{y} - \bar{\tilde{y}}\right)^2} \tag{82}$$

## 5.6 Variable Selection

The selection of the most informative variables in spectral data can decrease measurement time, increase model robustness, and allow for more straightforward interpretation [148], [149]. The available techniques to select those variables are itself a field of ongoing research [150]. The selection methods available are commonly clustered into the three groups visualized in Figure 26 [148], [150]. The filter methods perform the variable selection independently of the model, while wrapper methods include the model response in

| **Type** | **Name** | **References** |
|---|---|---|
| Filter | Variable Importance Projection (VIP) | [151] |
| Filter | Covariance Selection (CovSel) | [115], [152], [153] |
| Wrapper | Competitive Adaptive Reweighted Sampling (CARS) | [130] |
| Wrapper | Genetic Algorithm (GA) | [154] |
| Wrapper | Bootstraping Soft Shrinkage (BOSS) | [115], [155] |
| Wrapper | variable combination population analysis (VCPA) | [115], [156] |
| Wrapper | Uninformative Variable Elimination PLS (UVE-PLS) | [157] |
| Embedded | Soft-Threshold PLS (ST-PLS) | [158] |

Table 5: Overview of some commonly used variable selection methods in absorption spectroscopy.

their selection of variables. Finally, the embedded methods are a combination thereof with the variable selection incorporated directly into the machine learning model. Each method of selection offers its specific advantages and disadvantages. While wrapper methods are adjusted to the fitting algorithm at hand and thus have a higher chance of reaching the global optimum, they come at the cost of an increased computation time dependent on the model at hand. Filter methods, on the other hand, are inherently fast. Embedded methods are restricted to specialized fitting algorithms, which allow this intricate variable selection.

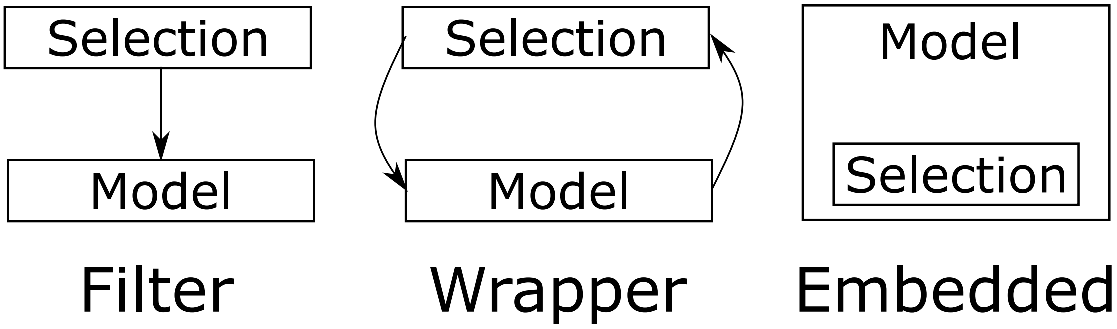


Figure 26: The three types of variable selection methods.

When handling spectral data, the high collinearity poses an additional challenge [149]. In addition, multiple target variables often need to be predicted simultaneously. An overview of standard variable selection methods used on spectral data is provided in Table 5. This overview highlights some important approaches in the field, as complete coverage is beyond the scope of this work. When selecting a method for a problem at hand, Mehmood et al. [149] pointed out the importance of the amount of data available and the properties of the data. Thus, no perfect method exists, but it needs to be chosen with the data and final applications in mind.

For the experiments in Section 7.1.5, basic Covariance Selection (CovSel) was employed [153]. CovSel is a stepwise approach that works independently of the model. It is closely inspired by the PLS algorithm. The first step selects the variable with the highest correlation to the predictions. Next, the data is orthogonally projected onto the selected variable. Those steps are repeated until the desired number of parameters is selected. This approach helps to reduce the effect of the high collinearity.

## 5.7 Spectral Machine Learning

Another application of machine learning in spectroscopy is the alteration of spectra. This branch of machine learning is difficult to cover completely, as it is considered a tool in multiple branches of spectroscopy. A traditional branch of spectral machine learning can be found in astrophysics, where MCR has long been used in telluric correction. The correction of spectral observations for the atmosphere the light had to penetrate on its way to earth [159]. Recent approaches also include AEs to disentangle the different spectral components of the incoming complete spectrum [160]. In the case of telluric correction, acquiring ground truth data is not always possible, making an exact evaluation and training very difficult. A similar problem is faced in the scatter correction approach for spectral preprocessing in IR absorption spectroscopy. Magnussen et al. have developed a CNN network to speed up scatter correction computations for infrared spectra [161]. A similar, extended network was later trained on simulated data to disentangle spectra of, for example, cells into the cell wall and cell interior spectral component and their refractive index [144].

### 5.7.1 Spectral Metrics

One main question when adapting spectra via machine learning is that of an expressive metric. Here, a broad range of metrics exist, extending from the metrics presented in Section 5.5.5. The most common metrics will be presented in the following, including their advantages and disadvantages concerning the fitting of spectra as the output data.

- **RMSE:** RMSE similar to MAE, MSE and Huber loss, which were previously presented, is easily implemented and is the most commonly used error function in spectral machine learning. It is easy to understand and evaluate. Nevertheless, it has some disadvantages when applied in spectral adaption. When handling large and small peaks simultaneously, the error is proportional to the signal height, thus putting more weight on the higher peaks. This can be countered by logarithmic preprocessing. The strictly vertical signal evaluation is another downside of RMSE. The surrounding data points are not considered, leading to unproportionally large errors at the steep slopes around peaks. Here, a small change in the peak location on the y-axis can lead to large errors on the x-axis. This puts a larger weight on the correct simulation of the slopes compared to the actual peaks.

- **Orthogonal Distance:** To include the x-axis information in the error computation, the orthogonal distance between a point and the closest point on the target spectrum can be computed. This computation is more complex than a simple RMSE. The x and y-axis in any spectral representation are of two different physical quantities; thus, a scaling factor is needed to weigh one towards the other. This adds an additional hyperparameter that is very sensitive. This leads to the metric being relatively unstable when training machine learning algorithms.

- **RIC:** Relative Integral Change (RIC) evaluates the area under the spectrum and has, for example, been used by Xue et al. [162]. Its computation is presented in Equation 83. Here, $\nu$ represents the wavenumber, while $y$ is the predicted value in its corresponding unit. Here, the difference of the area between the target and predicted spectrum in relation to the area under the target spectrum is used. This metric allows a better comparison between different heights of peaks and, therefore, molecules. Its computation is also more complex than that of the RMSE.

$$RIC = \frac{\int_{\nu_{min}}^{\nu_{max}} |\tilde{y}_i(\nu) - y_i(\nu)| d\nu}{\int_{\nu_{min}}^{\nu_{max}} \tilde{y}_i(\nu) d\nu} \tag{83}$$

- **Pearson / Spearman Correlation:** The use of correlation measures to compare different spectra has been investigated by Henschel et al. [163] to classify different spectra. They found Pearson correlation better to represent the most dominant features of a spectrum while Spearman correlation approximates band matches better [164]. Nevertheless, the correlation scores of spectra are difficult to interpret.

No perfect spectral metric exists, as each metric has advantages and disadvantages. It is advisable to consider a particular metric's problems before evaluating spectral machine learning. This difficulty also requires one to not only rely on the metrics available but also add a visual expert analysis, presenting the predicted, the target spectrum, and the residual. This allows the spectroscopist to analyze the overall quality of the prediction and, by zooming in on interesting regions, spot common problems in the prediction. Another solution is the evaluation of downstream problems, as can be seen in [161], where the output spectra were used as an input for further principal component and clustering analysis.

With this overview of the machine learning approaches and current state of the art in machine learning applied PAS, the models and results achieved in this thesis can be better interpreted. The next chapter will present the application of a deep learning model for spectral adaptation.

*Part III*

# Results

# 6 Spectral Adaption

Combining the spectral modeling capacities presented in Chapter 4 and deep learning techniques from Chapter 5, a way to adapt measured cross-sections to other environmental configurations has been developed.

This section describes the experiments and results to adapt Absorption Cross-Sections (ACSs) to different pressure configurations. The results presented in Section 6.3 have been published in [165]. In gas sensing, environmental conditions can greatly affect the exact shape and height of a molecule's ACS and thus each subsequent step in the sensing chain. While good simulations are available for smaller molecules, the interactions between the different atoms in larger molecules and many possible vibrational states disqualify this technique for larger molecular structures. As described in 4 for larger molecules, mainly measurement databases exist that do not cover all possible environmental configurations. The deep learning approach developed within this thesis is intended to require no additional chemical information on the molecule in question except a base absorption cross-section. This differentiates it from previous approaches, for example, Estimated-Pseudo Line Lists (EPLLs).

The approach in this chapter is visualized in Figure 27. The model takes an input spectrum at a specific pressure $p_1$ and outputs the same molecular spectrum at another pressure $p_2$. While the training is performed on simulated data, the evaluation and final application take place on measured data. Therefore, the model must bridge the domain gap between simulated and measured data.

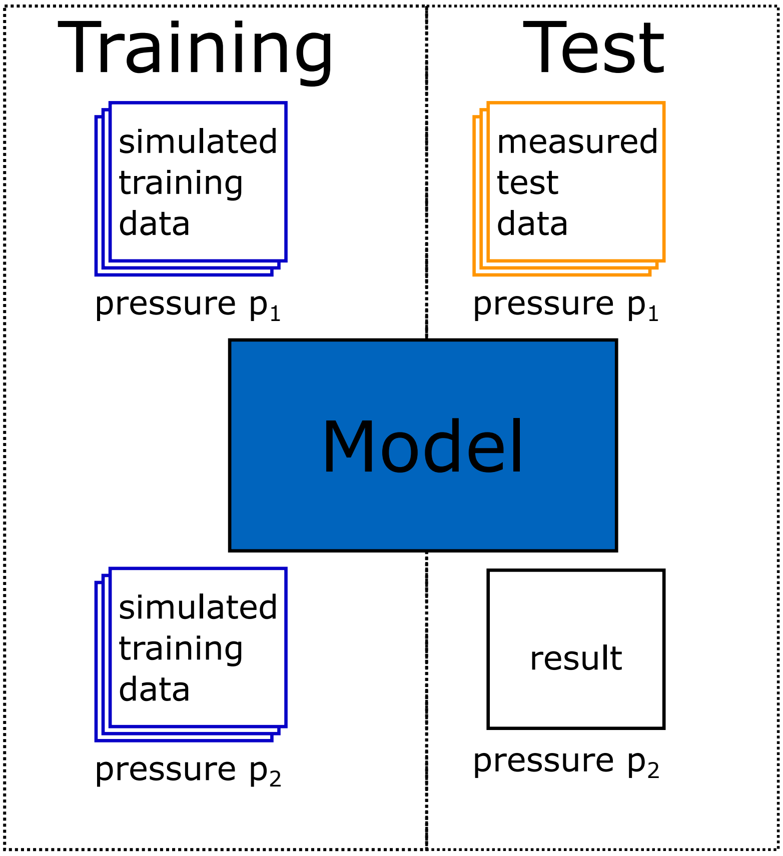


Figure 27: Schematic visualization of the training and testing paradigms. While during training the model only interacts with simulated data, at test time measured data is applied.

Simulated data for smaller molecules forms the basis of the training dataset for each model. The underlying assumption is that the model can learn the structure of the environmental influences from those molecules and transfer it to other, bigger molecules. This is shown to hold for pressure broadening, while for temperature changes and decreasing pressure, additional knowledge would need to be incorporated into the model to allow a correct adaption. First, the dataset creation for all following setups will be described in Section 6.1. The models applied for the experiments and their development are described in Section 6.2. Finally, the experiments and results on the pressure adaption to fixed values for increasing pressure are described in Section 6.3; the work described in this section has already been published in [165]. Additional results on successful continuous pressure adaption for pressure broadening and the problems for decreasing pressure adaption are described in Section 6.4. Finally, the difficulties regarding temperature adaption with this method are outlined, and directions for future research are given in Section 6.5.

## 6.1 Dataset Creation

The training data and validation data for all the following experiments were simulated using the High-Resolution Transmission Molecular Absorption Database (HITRAN) database [14] and the Voigt line shape simulation implemented in hapi [27]. The HITRAN database was chosen due to its coverage of many molecules and because it's the typical source for sensor development. While the Hartman-Tran line-shape simulation could provide more precise simulation results, it requires additional line-shape parameters that are only available for a few molecules and transitions. The simulation would revert to the Voigt simulation for all transitions without those parameters. We forgo this more precise simulation to retain more similarity throughout the dataset. For more information on the different simulation algorithms and line shapes, the interested reader is referred to Section 4. The out-of-dataset test data was acquired from the HITRAN cross-section database [78].

To separate this dataset into a training and validation partition, a split by molecule was considered. A leave-one-molecule-out evaluation for all samples was performed to create a validation set representing the structure of the entire dataset. Here, multiple molecules were identified that resulted in worse reconstruction Relative Integral Change (RIC) scores due to the structure of their absorption spectrum. From this evaluation, the more challenging molecules $H_2O$ were selected together with $OCS$ and $C_2H_4$. This selection is estimated to represent different molecular structures and spectral structures in the ACS.

To examine the models' performance on actual measurement data, which shows vast differences to the simulated data, the $\nu_2$ band of $\mathrm{ClONO_2}$ was selected. For $\mathrm{ClONO_2}$, measurements at a fine resolution and many temperature and pressure configurations are available. It was also chosen because a partition function and EPLL were readily available. The measured cross-sections by Wagner et al. [166] were chosen over previous measurements by Ballard et al. [79] as they are expected to be more accurate [79], [166]. The measured cross-sections were interpolated to a 0.01 $\mathrm{cm}^{-1}$ wavenumber resolution.

All simulated data parameters in the training and in-dataset test set were chosen to reflect best the out-of-dataset test measurements used for the final comparison. Air broadening was used for each simulation. A resolution of 0.01 $\mathrm{cm}^{-1}$ was chosen in a wavenumber range from 1200-2200 $\mathrm{cm}^{-1}$ corresponding to the measured absorption cross-sections. All other simulation parameters were also selected to correspond best to the external test data and are summarized for each dataset in Table 6.

| **Dataset** | Pressure fixed | Pressure continuous |
|---|---|---|
| **Wavenumber range** [$\mathrm{cm}^{-1}$] | 1200-2200 | 1200-2200 |
| **Resolution** [$\mathrm{cm}^{-1}$] | 0.1 | 0.1 |
| **Pressure** [Torr] | 19.6,30.1,71.6,115.7 | 19.6-115.7, 20 points |
| **Temperature** [K] | 219 | 219 |
| **Broadener** | air | air |

Table 6: Configurations for the spectral simulation used for dataset creation for the adaption experiments.

Due to the broad wavelength coverage and high resolution, each training sample would encompass 10000 input nodes, of which many did not contain any signal, especially for simpler molecules with fewer possible transitions in general. Each spectrum was separated into smaller snippets to increase the information in each data point and provide more training samples to the model. The exact size of those snippets found after hyperparameter tuning was 12 cm$^{-1}$. Each of those snippets had only 1 200 input nodes, and the snippets were screened to discard the samples that did not contain any signal and, thus, information. Each spectrum was separated into snippets of 12 cm$^{-1}$ with an overlap of 1 cm$^{-1}$ on each side. So the first snippet of each spectrum ranges from 1200 – 1212 cm$^{-1}$ and the second from 1202 – 1214 cm$^{-1}$. Each snippet is now screened if any signal higher than $10^{-21}$ cm$^2$/molecule is contained. This threshold was chosen as the model is intended to perform in the range from $10^{-16}$ – $10^{-23}$ cm$^2$/molecule, which is the typical range assessed with photoacoustic absorption spectroscopy. The threshold was chosen at $10^{-21}$ cm$^2$/molecule as this higher threshold ensures that slowly declining spectral wings from peaks are removed from the dataset as they do not contain useful information for the model. This procedure led to an overall number of 4 363 snippets for all simulated spectra. The distribution of spectra by molecule is presented in Table 7. A similar procedure was used to prepare the spectra for testing. In this case as well, snippets of 12 cm$^{-1}$ were created, but the overlap was extended to 3 cm$^{-1}$. To recombine the snippets, the outer 1.5 cm$^{-1}$ of each snippet side were discarded, as the model did show worse performance at the edges of the snippet.

As the spectral signal ranged over the vast range from $10^{-16}$ - $10^{-23}$ cm$^2$/molecule additional preprocessing was required to enable similar performance of the model on smaller and higher peaks. This is especially needed since the applied loss function (Root Mean Squared Error (RMSE)) behaves linearly and is thus directly proportional to the absolute cross-section value. Therefore, only large peaks will be modeled correctly without prior scaling, while the model neglects smaller ones. For this reason, a logarithmic scaling function has been designed. This function is presented in Equation 84 and scales the cross-sections from 5.96e$^{-23}$ and 1e$^{-15}$ cm$^2$/molecule between 0 and 1. As mentioned, the lower bound has been selected with a photoacoustic trace gas sensing system's limits in mind. At the same time, the upper limit is given by the data. To achieve this scaling capability, a base of eight was chosen. Higher bases are also possible but would increase errors at high peak amplitudes by downscaling those values additionally. A lower base, on the other hand, would increase the scaling range and thus include smaller amplitudes. The base of 8 was chosen as a compromise and optimization for this specific research.

$$s_{\text{scaled}} = \begin{cases} \log_8(s10^{15}) + 8/8, & \text{if } s10^{15} > 8^{-8} \\ s, & \text{otherwise} \end{cases} \tag{84}$$

| **Molecule** | **Number of snippets** |
|---|---|
| $C_2H_2$ | 129 |
| $C_2H_4$ | 103 |
| $C_2H_6$ | 194 |
| $CH_3Br$ | 243 |
| $CH_3Cl$ | 252 |
| $CH_4$ | 163 |
| CO | 145 |
| $COF_2$ | 205 |
| CS | 120 |
| $H_2CO$ | 143 |
| $H_2O$ | 512 |
| $H_2O_2$ | 124 |
| HCN | 129 |
| HCOOH | 116 |
| $HNO_3$ | 235 |
| $HO_2$ | 114 |
| HOCl | 81 |
| $N_2O$ | 131 |
| $NH_3$ | 340 |
| NO | 151 |
| $NO_2$ | 100 |
| $O_3$ | 76 |
| OCS | 210 |
| $PH_3$ | 110 |
| $SO_2$ | 102 |
| $SO_3$ | 81 |
| Total | 4363 |

Table 7: Number of extracted snippets per molecule used to create the in-dataset train and validation splits for the adaption experiments.

## 6.2 Model Architecture

Two main architectures were considered to adapt cross-sections to changing environmental configurations. A disentangled autoencoder allows for adaptions in the latent space and a change network. Those two approaches are schematically visualized in Figure 28. One difference between those two architectures is the compression of the data in the latent space in the Autoencoder (AE). In the change network, the information is never compressed in this form. In addition, the pressure adaption is applied differently to the two networks. While the AE is trained in an unsupervised or semi-supervised manner, i.e., the environmental adaption value is only integrated at the level of the latent space. This integration only takes place during inference and can be considered an adaption of the values in the latent space by addition or multiplication. On the other hand, the adaption value is always available with the input data for the change model. Therefore, the change network is a supervised model trained with the concrete adaption value available during training. In the following section, the AE approach will first be outlined and described, and experiment results will be presented. This includes the work of a master student, supervised by the author [167], and arguments why this approach was not further explored. Next, the change network will be presented, and improvements to the model will be discussed. Ending with the final architecture used for the Sections 6.3 to 6.5.

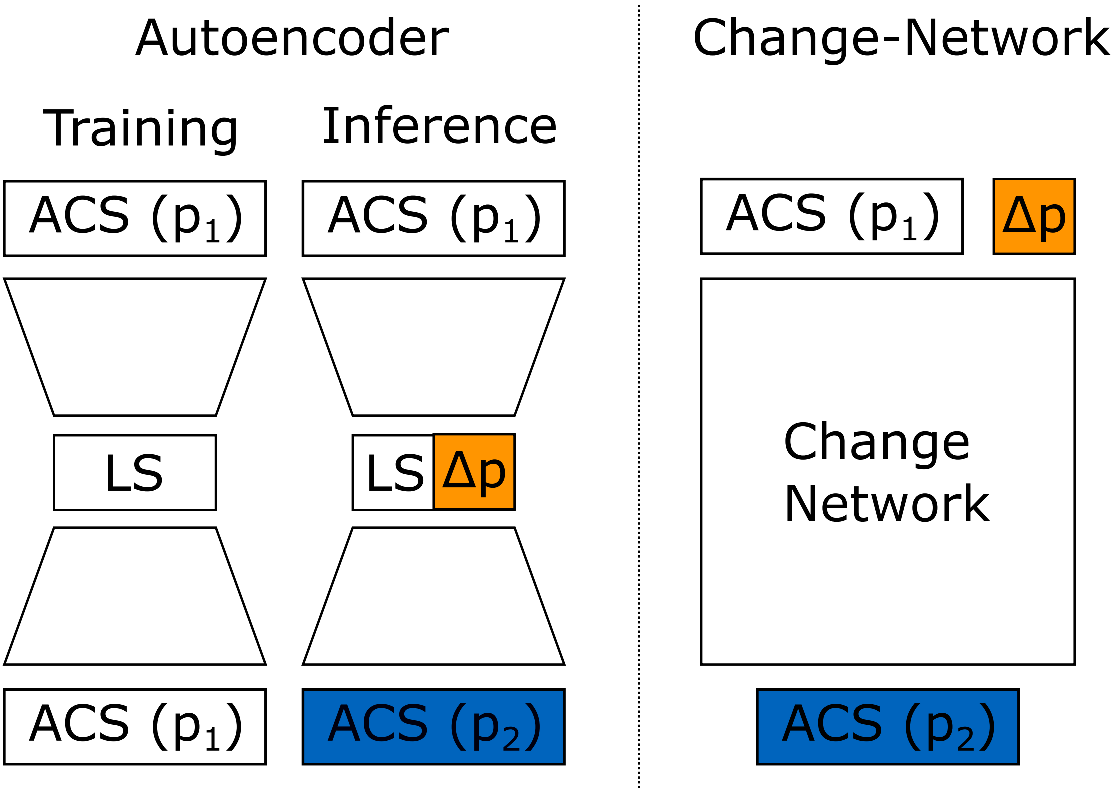


Figure 28: Schematic visualization of the AE and change network. During training the autoencoder reconstructs the ACS at the input pressure configuration. Only during inference, the latent space is adapted, which leads to the reconstruction of the ACS at a different pressure configuration. The change network on the other hand receives the pressure change variable as an input and directly predicts the ACS at the requested output pressure configuration. The integration of the pressure change variable is indicated in orange.

### 6.2.1 Autoencoder

For a first evaluation of the ability of AEs to recover the spectrum from a latent space, a simple convolutional AE was constructed and trained on the simulated ACSs. Differently from later models, it worked on snippets of a spectral range of 24cm$^{-1}$ and still employed min-max scaling for the spectral values, which was later found to discourage good reconstruction on the smaller peaks in the ACSs. The layout of the AE is visualized in Figure 29. All hyperparameters, including the number of filters and kernel sizes, were trained using Bayesian hyperparameter tuning as implemented in weights and biases for 65 epochs. The final model was trained using the Stochastic Gradient Descend (SGD) optimizer and the hyperparameters supplied in Table 8. It had a mean validation RMSE of 0.0500, resulting in decent reconstructions.

| **Hyperparameter** | **Value** |
|---|---|
| Optimizer | SGD |
| Learning Rate | 2.895e$^{-4}$ |
| Weight Decay | 2.613e$^{-5}$ |
| Number of Epochs | 150 |
| Kernel Sizes | 3, 10, 8 |
| Number of Filters | 176, 100, 34 |
| Latent Size | 14 |

Table 8: Hyperparameters of the simple convolutional autoencoder after hyperparameter tuning.

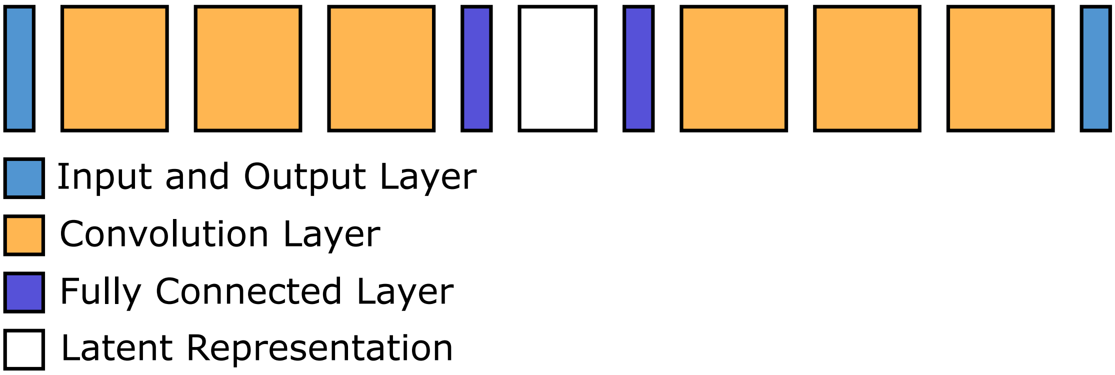


Figure 29: Model architecture of the simple convolutional autoencoder. The encoder and decoder are convolutional layers, with additional fully connected layers interfacing the latent representation.

To evaluate the influence of pressure on the latent space, the encoder was run for each molecule and different pressure configurations. The latent encodings were saved and averaged over all molecules. The results are visualized in Figure 30. The plot shows an inverse dependence of the three upmost variables on the pressure. Therefore, a change in pressure is encoded in the latent space of this model. A similar trend was visible in three independent runs. This leads to the conclusion that a good disentanglement could enable the differentiation of the pressure component in the latent space and its adaption to create reconstructed spectra with a different pressure configuration.

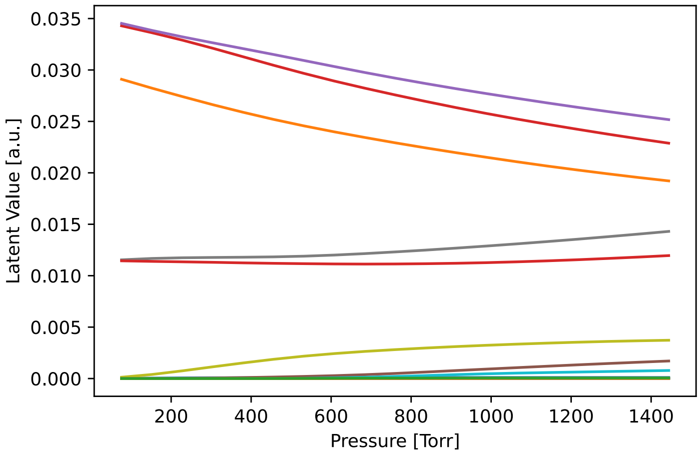


Figure 30: Mean of all molecules latent variables by pressure for the simple convolutional autoencoder. A trend in some latent variables with the pressure becomes visible (purple, red, orange). Others are not influenced by the pressure change (green, turquoise).

Motivated by those results, a further investigation of disentangling semi-supervised AEs, namely a Semi-Supervised Adversarial Autoencoder, was performed by Tristan Schaar under the author's supervision [167] based on the work of Makhzani et al. [168]. Disentanglement refers to the separation of the latent representation into different parts. The technique of a Adversarial Autoencoder (AAE) uses adversarial training to create this separation in the latent representation. The aim here is to separate a part of the latent space, which is highly correlated with the pressure change, while the remainder of the latent space is not. Semi-supervised training was used to increase the correspondence of the separated part of the latent representation and the pressure values by using the pressure value as a supervised learning target. For a more in-depth explanation of the approach, the interested reader is referred to the original work [167]. The results of this work and further interpretation will be presented in the following. Schaar found that while the relative pressure change could be seen in the disentangled latent variable, it was not stable over multiple training runs and sometimes not even over all molecules. In addition, no network was able to predict the pressure from the input spectrum reliably. Nevertheless, reconstructions were, except for minor offset errors, very reliable. The complete results can be found in the original work [167]. Through this work, the disadvantages of the AE approach became eminent:

- **No control of absolute values:**
  Due to the relative relationship of the latent representation and the pressure value, no absolute value could be set for the resulting spectrum pressure even after disentanglement.

- **High model dependence:**
  Schaar noted a strong dependence of the disentangled variable on each model training. Even if training parameters were kept constant, the unsupervised part of the training results in stark differences between the models when initialized from random weights.

- **Insufficient reliability:**
  Even if a good relative correlation between the latent variable and the pressure was found, this correlation was not reliable as it sometimes reverted for certain molecules or snippets.

This insufficient performance of the network can be attributed to missing information for the network to perform the task in question correctly. Even though more information has been provided to the model by encoding the molecule in question to dissolve this concern, this has not been sufficient. Especially the actual position on the wavenumber range was still unknown to the model. In addition, it would need to be informed about the molecular structure, either by the transition lines as used in HITRAN or another molecular information type as can be encoded in molecular fingerprints [169]. To the author's knowledge, simple molecular fingerprint structures focused on the rovibrational states of molecules were not available to perform this task. In addition, the unsupervised training manner introduced a broader variation in the results and would have required a much larger amount of data to disentangle the variable in question successfully. Therefore, AEs were not further pursued in this work, and more focus was put on the change networks, a supervised learning-based approach.

### 6.2.2 Change Network

For the reasons presented in the previous section, all subsequent work focused on the change network. The training paradigm of the change network differs strongly from an AE. Most notably, it is a fully supervised training, where the pressure difference and the shape of the target output spectrum are known during training. The change network was constructed carefully, starting from the simplest form and introducing additional complexity only where necessary. The final structure presented in the later part of this section is the result of many prior experiments. The findings leading to the final setup are summarized in the following:

- **Predicting the difference spectrum vs. the resulting spectrum:**
  The two concepts of the prediction of the difference spectrum between two environmental configurations versus the prediction of the resulting spectrum at another environmental configuration are depicted in Figure 31. In this case, the advantages of predicting the resulting spectrum over the difference spectrum prevailed. Those advantages include retaining the input structure through later layers and a higher local awareness. This allowed those networks to apply more complex dependencies on the original structure. The prediction of the difference spectra at different pressure configurations did not yield good results overall in initial experiments. This is probably due to the additional task of the network, to remove all contributions of the input spectrum from the data while simultaneously computing the correct difference spectrum. Those experiments were conducted with simple convolutional and fully connected networks.

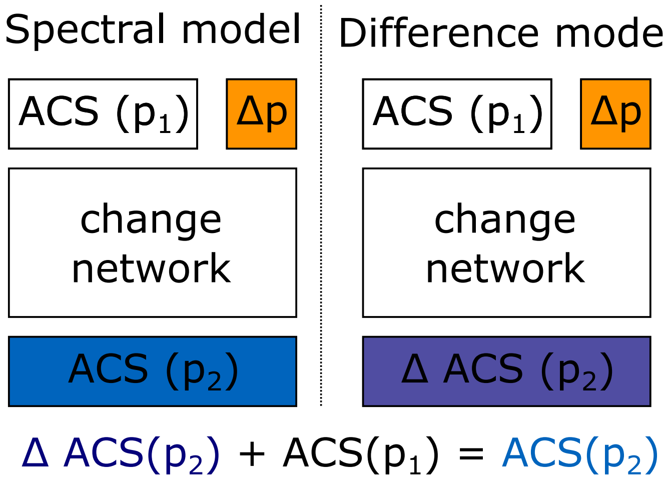


Figure 31: Schematic of the prediction of the difference spectrum vs. the prediction of the resulting spectrum for changing environmental configurations. The spectral model directly outputs the ACS at the required pressure configuration ($p_2$). The difference model output can be added to the input ACS ($p_1$) to compute the ACS at the required pressure configuration ($p_2$).

- **Fully connected vs. convolutional neural networks:**
  Fully connected networks were used for the first evaluations as the simplest form of Artificial Neural Networks (ANNs). But those models quickly converged to a local minimum, which was the mean of all training spectra. Therefore, convolutional layers were introduced. Even a simple three-layer convolutional neural network, similar to the ones used as the encoder of the AE approach, was able to produce spectral shapes as output. Therefore, convolutional layers were retained for the remainder of this work.

- **Fixed pressure prediction:**
  To ease the task for the network and circumvent initial troubles with integrating the delta pressure value into the network, separate networks were first trained for each delta pressure. This increases the computational demand for the final use but was found to perform better. Continuous pressure-changing networks will be reintroduced in Section 6.4.

- **Residual connections:**
  To further improve the performance of the Convolutional Neural Network (CNN) residual blocks, as introduced in [111] and described in Section 5.3 were applied to the network. This approach is known to counter the vanishing gradient problem and allow a deeper network structure. In the case of this application, it could also allow the model to focus on different aspects of spectral change in each residual block. An example residual block, as implemented in this work, is depicted in Figure 32 and contains two convolutional layers and a scaling component. The introduction of residual blocks improved the models' performance and was retained for further experimentation.

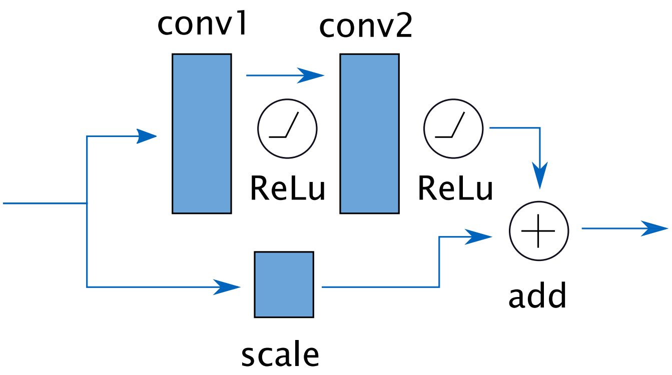


Figure 32: Schematic of the residual block used in this work. The input is split into two paths, one passing through two convolutional layers while the other is scaled and added to the output of the first path before leaving the residual block.

- **Batchnorm layers:**
  Batch normalization is a commonly used tool in the construction and training of ANNs. It introduces an additional layer before the convolutional layer, which takes an input batch and transforms this batch to zero mean and unit variance. Even though they are frequently used, their effect is still only partially understood [170], [171]. Introducing batch norm layers in the CNN model reduced the models' overall reconstruction performance. This is attributed to the fact that the absolute output value is highly dependent on the input spectrum at that position. The expected output spectrum is very similar in absolute values to the input spectrum. This dependence is destroyed through the introduction of batch norm layers, rendering the task more difficult for the network as it has to counteract batch normalization. Therefore, no batch normalization was employed in the final model.

- **Very large vs. small convolutional kernels:**
  The kernel size of convolutional filters is one of the hyperparameters to be tuned. While commonly small kernel sizes of three to maybe twelve are chosen, some works in spectroscopy suggested using very large kernel sizes of 20-90 [98], [172]. To evaluate this notion of larger kernel sizes, a project work and master thesis was performed by Nikhitha Gudur and supervised by the author [173]. The hypothesis of a performance gain through larger kernel sizes was tested on two Infrared (IR) datasets, the mango spectral dataset [174] as well as the melamine dataset [175]. Both experiments could not verify this hypothesis. In fact, better results were achieved using smaller kernel sizes. Similarly, a short experiment employing larger kernel sizes (20-90) was tested on the spectral adaption at hand by the author but was also not found to improve the overall performance. Therefore, smaller kernel sizes in the range of three to twelve were employed for all future experiments.

- **Loss evaluation:**
  As mentioned in Section 5.7.1, the metric to be optimized by the network plays a crucial role in the result. Therefore, different losses were evaluated for the change network. The overall metric for the final evaluation was chosen as the RIC as this metric best covers the small and larger peaks. The four

loss types that were compared in performance are RMSE, orthogonal loss, Pearson correlation, and Huber loss. The results of ten independent runs with the same architecture per loss are visualized in Figure 33. Using the Pearson correlation as a loss metric did not lead to stable training, as can be seen from the large standard distribution, and was therefore discarded. The mean performance of the orthogonal loss (RIC = 0.56) was slightly worse than that of Huber and RMSE losses (RIC = 0.54). Therefore, the RMSE loss was selected for further experiments, resulting in one less hyperparameter during training.

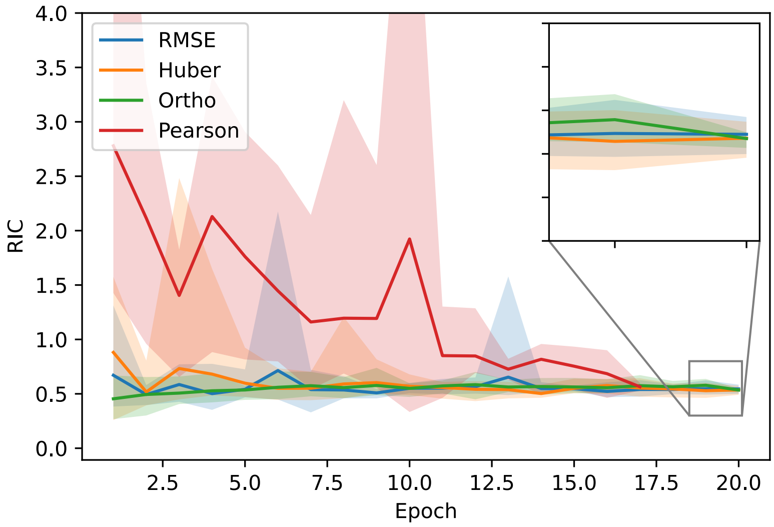


Figure 33: Mean RIC score of ten independent runs for different optimization metrics.

- **Snippet size:**
  The size of the snippet introduced into the network was also tuned during the experiments. The possible values were set to 240, 300, 600, 1200, 1700, and 2400. The edges of the snippets were adapted respectively to 20, 50, 50, 100, 100, and 200. Those were not included in the validation. The results of 16 independent runs are provided in Table 9. While the best RIC was achieved with a small snippet size of 300, visual inspection revealed an insufficient reconstruction performance of the smaller snippet sizes (240, 300, 600). Due to this incoherent setting of the small snippets, a snippet size of 1200 was chosen for all subsequent experiments.

- **Squeeze and Excite structure:**
  The Squeeze and Excite (SE) block, as introduced by Hu et al. [112], was already presented in Section 5.3. Those blocks were introduced into the model after each residual block to improve reconstruction performance. They were implemented as visualized in Figure 34. And contain a global average pooling layer and two fully connected layers. The output is scaled with a sigmoid function to range between zero and one. They also provide an additional possibility to introduce

| Snippet Size | Average RIC |
|---|---|
| 240 | 0.58 |
| 300 | **0.54** |
| 600 | 0.56 |
| 1200 | 0.57 |
| 1700 | 0.61 |
| 2400 | 0.59 |

Table 9: Average RIC score for different snippet sizes from 16 independent runs per size.

a delta pressure value into the network, as will be further explained in the following bullet point. The increase in performance of the SE block was evaluated for the inclusion of an SE block after every residual block. While the models including more SE blocks took a longer training time to converge, the overall RIC after training improved with the number of SE blocks. Therefore, each residual block had a concatenated SE block in the final model.

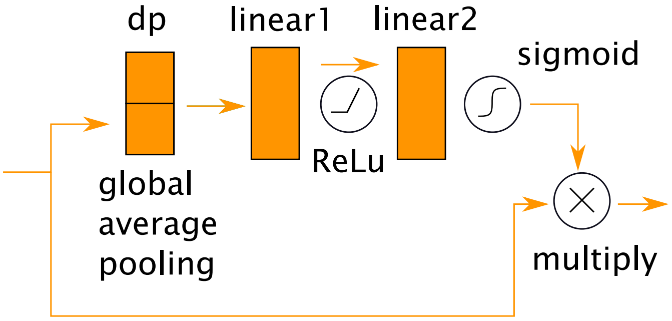


Figure 34: Schematic of the Squeeze and Excite block introduced after the residual blocks. The input is split into two paths. The upper one passes through a global average pooling to extract the channel means. Those are extended with the delta pressure and fed to a two-layer linear network. The output is scaled by a sigmoid function and multiplied by the input data from the second path.

- **Introduction of the delta pressure value:**
  While up to now, only one set pressure change was evaluated per model (i.e., 760 Torr - 912 Torr), for the continuous models, the delta pressure value needs to be provided to the model at multiple locations. A concatenation to the input spectrum was not performed due to the convolutional kernels, which would not consider the absolute value but the shape this introduced with the last value of the spectrum. First, a scaling of the convolutional output was considered but not found to perform well. Finally, the introduction inside the SE layers was performed as presented in Figure 34. This did not disturb the performance in the one delta pressure models and allowed for a more considerate integration of this delta value for the continuous models. Those will be separately evaluated in Section 6.4.

This leads to the final architecture presented in Figure 35. The network input is the preprocessed ACS at the base pressure. This corresponds to an intensity vector with 1 200 input features. The network

| **Parameter name** | **Value** |
|---|---|
| Kernel size block 1 | 5 |
| Kernel size block 2 | 7 |
| Kernel size block 3 | 11 |
| Kernel size final convolution | 5 |
| Loss function | RMSE |
| Optimizer | Adam |
| Learning rate | 1e-2 |
| Momentum | 0.89 |
| Squared momentum | 0.91 |
| Weight decay | 5e-6 |

Table 10: Hyperparameter settings used throughout the spectral adaption experiments.

has three residual blocks (Res1, Res2, and Res3). Each residual block contains two one-dimensional convolutional layers (conv1 and conv2). The channel size was kept constant at 64 for all residual blocks. A ReLu was chosen as a non-linearity, and the skip connection of the residual block included a scaling component. Each residual block is followed by a SE block. Inside the SE block, global average scaling is used to reduce each of the 64 convolutional channels to one value. The delta pressure value, scaled between zero and one, is concatenated to the average values. Since the current models only predict one preset pressure change, this delta pressure component corresponds to an additional constant bias term. The concatenated values are followed by two linear layers with 64 neurons each. While the linear1 layer has a ReLu activation function, the linear2 layer employs a sigmoid activation function. This is chosen as advised by [112] as the sigmoid ensures the output values are between zero and one, which is correct for scaling. Those scaling values are finally multiplied with their respective channels. Finally, the model's hyperparameters were tuned with Bayesian hyperparameter tuning in weights and biases for 65 epochs. This leads to the training configuration provided in Table 10. Even though a new set of hyperparameters could improve the performance of the continuous models, the same configuration was used due to time constraints. Nevertheless, this configuration was sufficient to show the applicability and problems of those models.

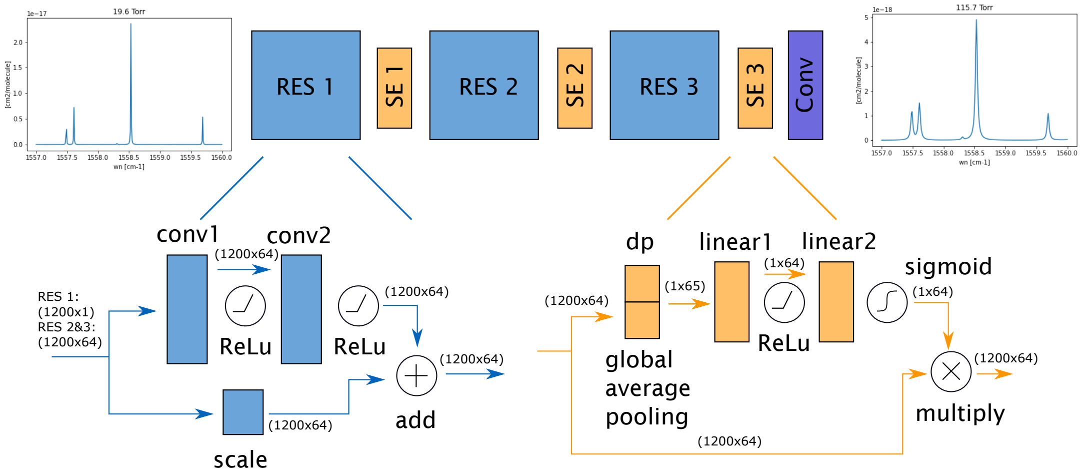


Figure 35: Overall architecture used for spectral adaption. The input is the ACS at base pressure. The network is constructed of three residual blocks followed each by a squeeze and excite block. The output is generated by a final convolutional layer. The data shape is indicated for the residual and the squeeze and excite blocks (adapted from [165]).

## 6.3 Pressure Adaption to Fixed Values

For pressure adaption, only increasing pressure settings were evaluated. This restriction was made to ensure a high prediction quality. In the case of a pressure decrease, additional structures might become visible in the spectrum that were previously merged. This is exemplarily visualized in Figure 36. In this ACS simulated for water, the finer structures between $1742\text{cm}^{-1}$ and $1750\text{ cm}^{-1}$ only show at the lower pressure configuration. The reconstruction of those merged features solely from the higher pressure spectrum is impossible without additional molecular structure information. As the proposed network only considers the base spectrum and the requested delta, no correct prediction can be made. The exact pressure difference where those additional features emerge differs for each molecule and pressure setting. Therefore, pressure decrease was completely disregarded for the following experiments. Possible ideas and approaches to include additional information into the model that might enable this prediction will be outlined and explained in Section 6.5.

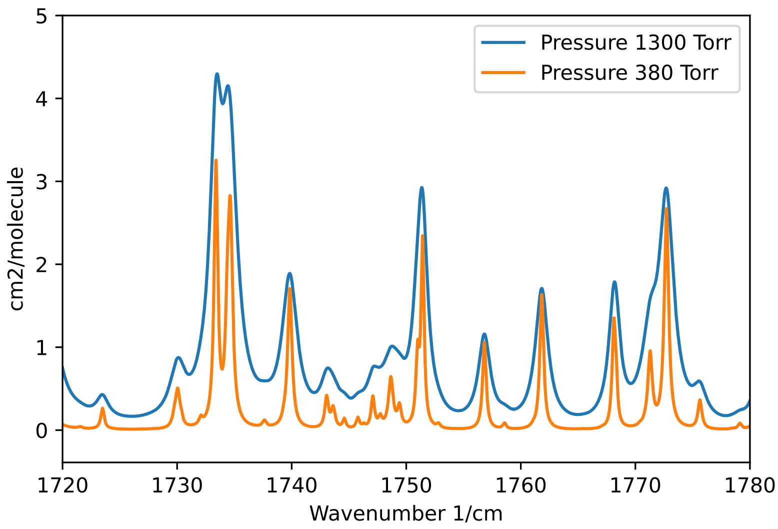


Figure 36: Absorption cross-section of water at two different pressure settings. For a higher pressure the broadening of the absorption structures can be observed.

First, an in-dataset evaluation was performed on the pressure adaption for fixed values. The in-dataset indicates that the validation cases here stem from the same domain, simulated cross-sections. The validation dataset selected consists of OCS, $H_2O$ and $C_2H_4$. No cross-section snippets from those molecules were used during training, nor hyperparameter tuning. Fixed pressure adaption refers to training a separate model for each pressure change, so in this case, three different models were trained to predict the spectrum at 30.1, 71.6, and 115.7 Torr from the 19.6 Torr base cross-section. Secondly, the models' performance on a measured ACS of a larger molecule, namely $ClONO_2$, was evaluated. Here, the same setup was used, but the task for the model was more challenging, as the measured ACSs contain measurement artifacts.

### 6.3.1 In-Dataset Evaluation

The different scores and a visual inspection were combined to evaluate the models' performance on the simulated validation set. The RMSE and RIC scores and the maximum residual were computed over the entire spectral range. So, even though the model was applied to the smaller snippets, those were concatenated, excluding the overlapping edges, and the scores were computed over the entire spectral range. The resulting metrics from four independent runs for the training and validation set are presented in Table 11. A lower score indicates better performance.

A first insight from those metrics is that overfitting did not occur, as the training RMSE scores are even slightly greater than validation scores. In the case of overfitting of the network, the validation scores would be worse / higher. Overall, the RIC score of less than 3% for all samples can be considered very good. The maximum residual is below 7% for all samples, indicating good similarity over the entire

| **Metric** | 20.5 Torr | |
|---|---|---|
| | Train | Validation |
| RMSE | 4.268e-21 | 2.995e-21 |
| RIC | 0.0264 +- 0.0010 | 0.0207 +- 0.0012 |
| Max Residual | 0.0338 | 0.0528 |

| **Metric** | 52.0 Torr | |
|---|---|---|
| | Train | Validation |
| RMSE | 4.522e-21 | 4.102e-21 |
| RIC | 0.0292 +- 0.0017 | 0.0256 ± 0.0020 |
| Max Residual | 0.0575 | 0.0685 |

| **Metric** | 96.1 Torr | |
|---|---|---|
| | Train | Validation |
| RMSE | 4.191e-21 | 4.075e-21 |
| RIC | 0.0288 ± 0.0011 | 0.0264+-0.0008 |
| Max Residual | 0.0533 | 0.0688 |

Table 11: Mean validation and train metrics from four independent runs. Computed on the full cross-section per molecule. (adapted from [165])

spectral range. This shows that at no position, an overly high outlier is generated. The validation RIC and maximum residual are slightly increasing for higher pressure differences, which was expected as the difference between the input and output spectrum also increased. But, even though the prediction for the smallest delta (20.5 Torr at 30.1 Torr) shows better scores than the other two, no linear trend can be observed. Thus, a higher pressure difference does not necessarily result in worse performance. The standard deviation of the four models computed on the RIC score is small, which indicates a stable method.

Figure 37 presents the residual for each validation spectrum. On the bottom, the base cross-section is shown. Above it, the residual at each spectral position is presented for 115.7, 71.6, and 30.1 Torr from bottom to top. The residual was computed as $(target - prediction)/max(target)$. From the plots, the location of the highest residuals also shows to be located at the peak position. To better illustrate this effect in Figure 38, a zoomed-in cross-section of $H_2O$ is presented. Here, the difference at the peak is highlighted, and it can be seen that the residual seems proportional to the peak size. This was seen throughout the validation set and can be attributed to the RMSE error and the effects of logarithmic scaling, where errors on the higher values are down-scaled. The low standard variation shows the robustness of this training method.

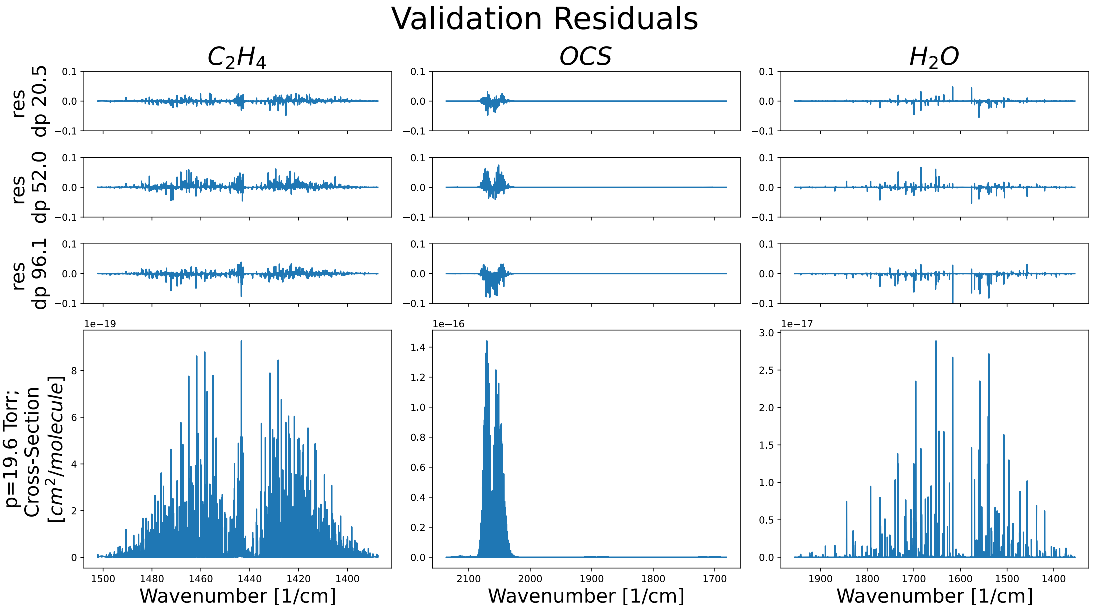


Figure 37: Base absorption cross-section at the bottom and the residual computed for the three delta pressure settings for the three in-dataset validation molecules. (taken from [165])

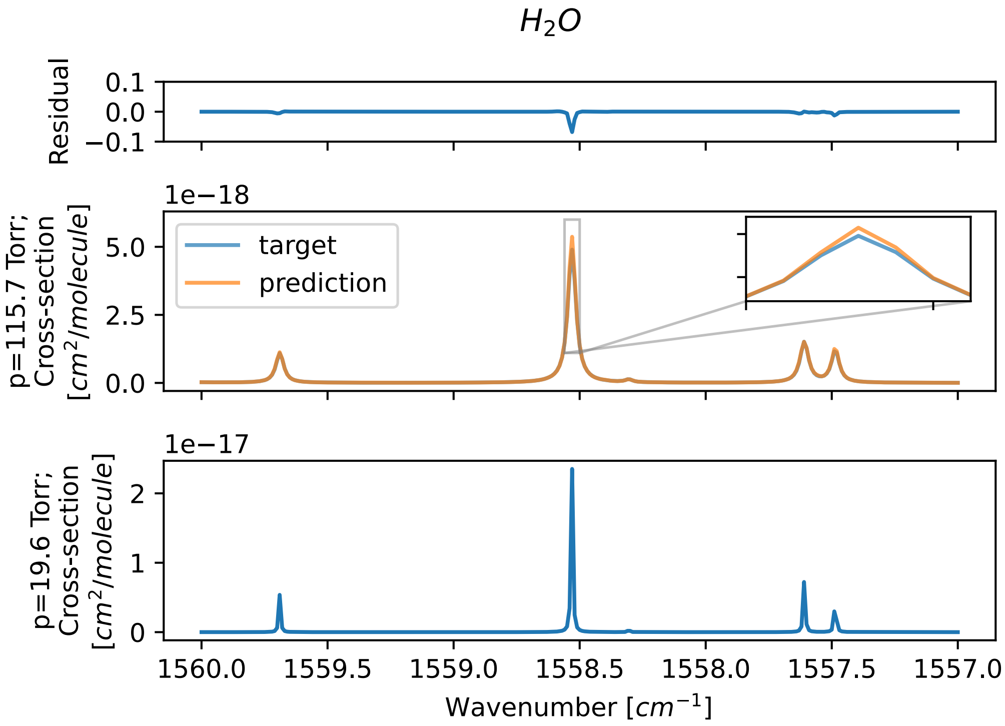


Figure 38: Visualization of the typical error position. The bottom plot shows the base absorption cross-section, the middle plot the target, and the actual prediction of the model with a zoom-in on the peak position. The upmost plot shows the residual of the prediction. (taken from [165])

### 6.3.2 Out-of-Dataset Evalutation

This overall good performance was next evaluated on an out-of-dataset test setup. The overarching goal of the training is to use the system to change an ACSs for which simulation is not yet possible. The measured ACSs typically belong to larger molecules, which cannot be modeled correctly due to the many interactions of their molecules and groups, which cause a vast number of rovibrational states. For those molecules, ACSs estimations are more complex, and line databases like HITRAN do not contain parameters for their simulation. Therefore, typically, measured databases are used as estimates. If a measurement at a specific environmental configuration is unavailable in the database, an estimation through EPLLs can still be performed. Those EPLLs are generated from multiple high-precision measurements. But as for any measurement, they often show artifacts such as channel fringes, residuals from the instrument function, or small traces of other gases [79], [166]. Their generation is described in Section 4.6.

The same four independent models as in the in-dataset evaluation were used for the out-of-dataset evaluation. The molecule $ClONO_2$ was chosen for the evaluation since the corresponding partition function for $ClONO_2$ is known, and the EPLL is available. The EPLL of $ClONO_2$ was computed for each of the pressure configurations to compare the two methods directly. For the EPLL computation, the pseudo line data was downloaded from the Jet Propulsion Laboratory homepage [81] and the partition function by Gamache et al. [30] were used in combination with hapi [27].

The resulting cross-section predictions and targets are visualized in Figure 39 to 41. The bottom plot visualizes the deep-learning model prediction and the target cross-section. The two top plots show the residual of the EPLL (middle plot) and the deep-learning prediction (top plot). From the residual plot, a slightly better performance of the EPLL can already be seen. This also shows when the RIC scores of the entire cross-section are compared in Table 12. The overall RIC scores of the deep learning model are around 5%, which is acceptable but a decrease from the in-dataset evaluation. The EPLL RIC scores are between 2 and 3%.

The better performance of EPLLs on the prediction is expected, as they contain more information about the molecule in question. As explained in Section 4.6, EPLLs are founded on a wide range of measurements of a molecule, in this case including the same ones as used for evaluation. Therefore, they perform much better in suppressing the instrument artifacts, which are still visible in the deep learning setup. This can be seen especially in the sine-shaped prediction from 1320-1310 $cm^{-1}$ visible in each prediction from Figure 39 to 41. This shape is eminent in the measured base spectrum at 19.6 Torr and can probably be attributed to water residue. It is corrected for in the EPLL and the measurements at 40.1 and 71.6 Torr, but a similar structure can be seen in the original 115.7 Torr measurement. This finding highlights the dependence of the deep learning adaption on the used base spectrum. Any error in this base spectrum will be propagated to the adapted spectra. For this reason and the other domain differences between measured and simulated data, the RIC scores of the deep learning models are around 2-3% higher than for the in-dataset evaluation. Notably, the standard deviation of the deep learning models is increased for the out-of-dataset validation. This might be reduced if more validation data is used in the out-of-dataset evaluation. Therefore, model inspection and selection are critical for future trials.

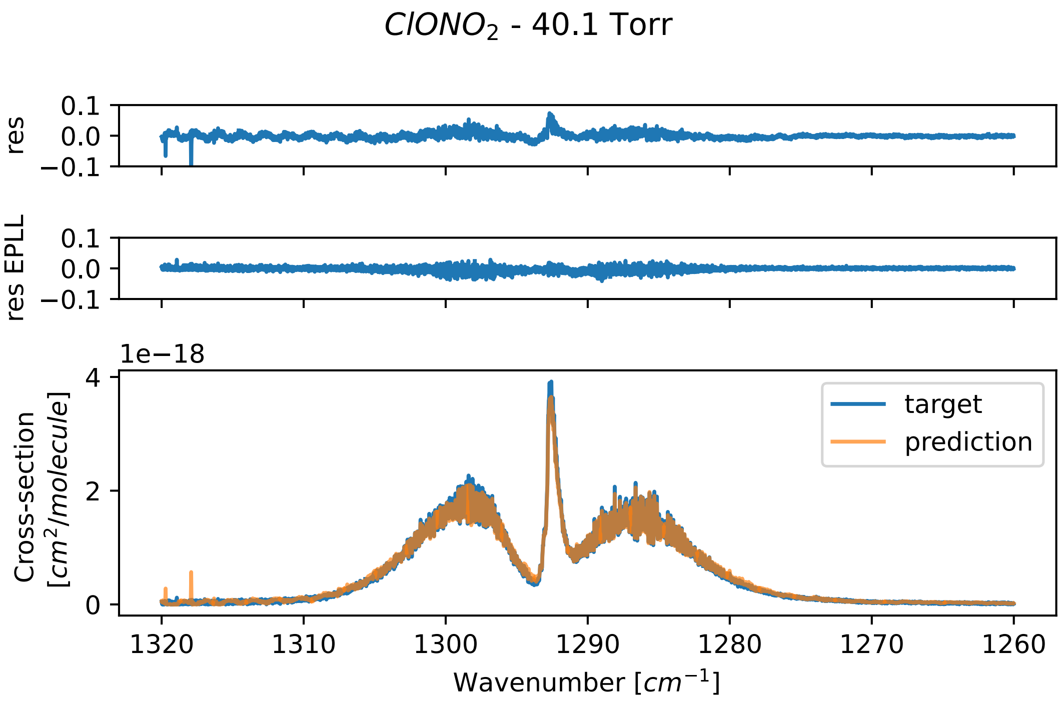


Figure 39: Validation results for $ClONO_2$ at 40.1 Torr. The bottom plot shows the measured/target and predicted cross-section. The two upper plots show the residuals. The middle plot shows the residual from measurement and EPLL simulation (res EPLL). The upmost plot shows the residual (res) between measurement and prediction (res). (taken from [165])

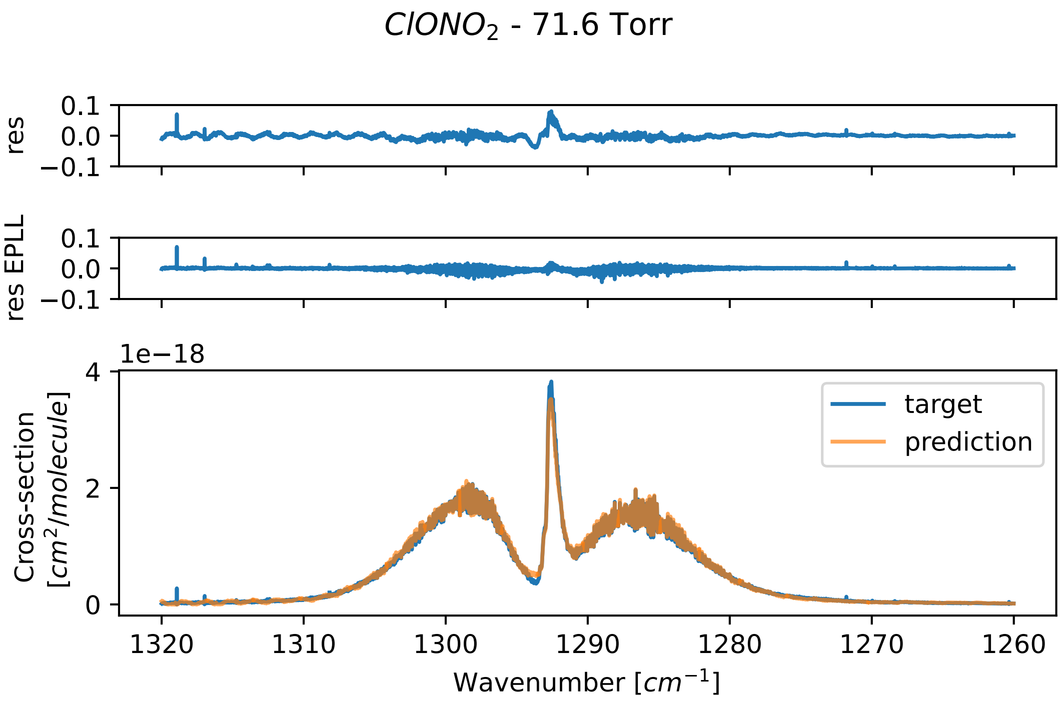


Figure 40: Validation results for $ClONO_2$ at 71.6 Torr. The bottom plot shows the measured/target and predicted cross-section. The two upper plots show the residuals. The middle plot shows the residual from measurement and EPLL simulation (res EPLL). The upmost plot shows the residual (res) between measurement and prediction (res). (taken from [165])

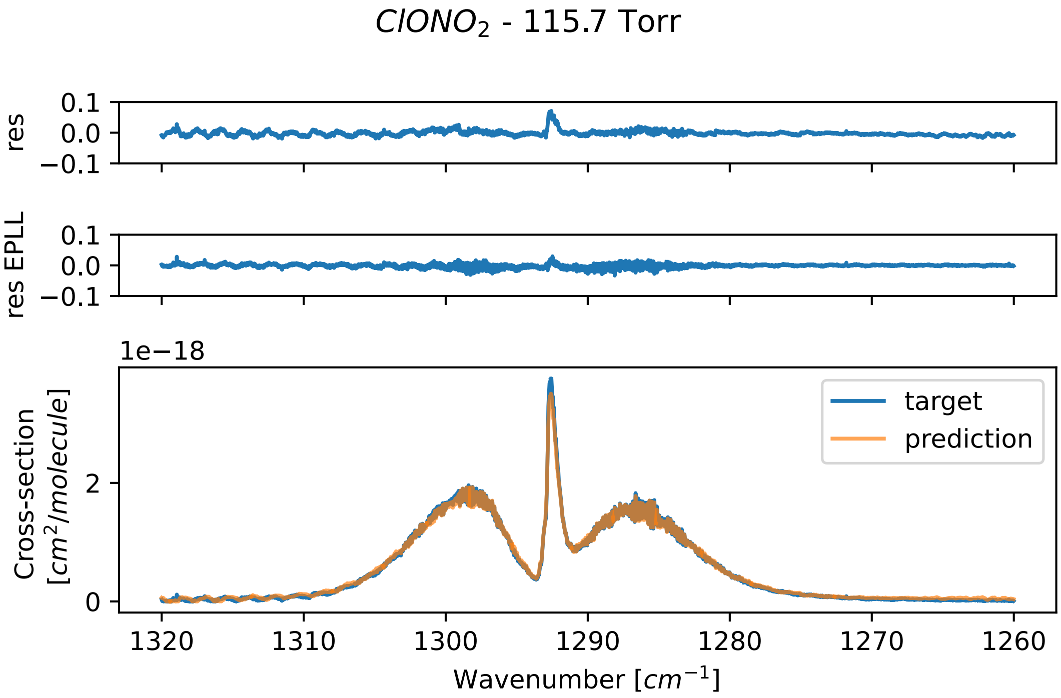


Figure 41: Validation results for $ClONO_2$ at 115.7 Torr. The bottom plot shows the measured/target and predicted cross-section. The two upper plots show the residuals. The middle plot shows the residual from measurement and EPLL simulation (res EPLL). The upmost plot shows the residual (res) between measurement and prediction (res). (taken from [165])

| **Target Pressure** | **RIC EPLL** | **RIC Model** |
|---|---|---|
| 40.1 Torr | 0.0320 | 0.0520 +-0.017 |
| 71.6 Torr | 0.0185 | 0.0490 +- 0.022 |
| 115.7 Torr | 0.0235 | 0.0520 +- 0.021 |

Table 12: RIC scores on $ClONO_2$ for the EPLL and the deep learning model. (taken from [165])

## 6.4 Continuous Pressure Adaption

As training a separate model for each pressure setting requires additional time and resources if multiple settings are under investigation, a combined, continuous model was implemented and evaluated. For this setup, the training was slightly adapted from the previous setting. Again, the base cross-section remained at a constant low pressure while the model was trained on different delta pressures, which were all increased compared to the base pressure. The temperature of the ACS was kept constant. The same simulation settings as for the previous model were chosen to make the models better comparable. The hyperparameters were not adapted from the fixed model training. The delta pressure value was introduced to the model as a concatenation inside the SE block as described in Section 6.2. The full grid of 20 equidistant points from 19.6 to 115.7 Torr was used during training. The delta selected for each training spectrum was chosen randomly.

The model was evaluated separately for each delta pressure setting, and additional delta pressure values were introduced in the test set to ensure the model interpolated correctly between the training set points. For this, 60 equidistant points between 19.6 and 115.7 Torr were created. Three independent models were trained to evaluate the model deviation. As the resulting RIC depends on the delta pressure setting of the evaluation, it is depicted in Figure 42. The scores were computed on the in-dataset validation set consisting of the molecules $H_2O$, OCS, and $C_2H_2$ as for the fixed pressure models. The orange line shows the model prediction at the pressure settings used for training and their standard deviation. The green line shows the model's prediction at additional pressure settings that were never before seen by the model and proves that the model can interpolate to unseen pressure configurations. The fixed pressure model's results are depicted in blue with error bars. For the continuous model, a higher overall RIC becomes evident, which linearly increases with the delta pressure setting. The increase with a higher pressure setting was expected, as the change in the cross-section is proportional to the delta pressure setting. In addition, the standard deviation increases with delta pressure. The mean RIC increases from around four to about nine % over the range of almost 100 Torr.

A higher error was expected compared to the fixed models, as the task is now more complex. The difference could be decreased by hyperparameter tuning adapted to the new task, but it is not likely to reach the same level. Therefore, a continuous adaption model can be used if less accuracy in the prediction is required, but fixed pressure models remain important for increased accuracy.

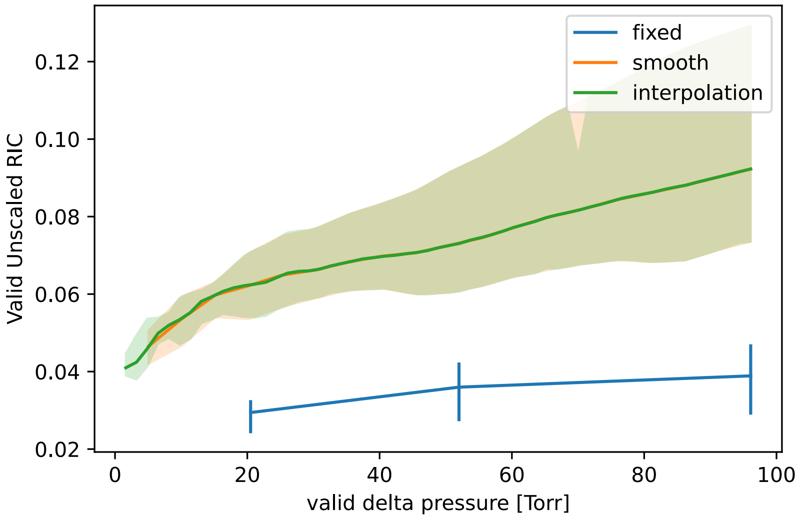


Figure 42: In-dataset validation RIC for the continuous pressure adaption model compared to the fixed model. While the fixed model shows got interpolation (green curve is similar to orange curve), the fixed networks (blue) show a lower overall score.

## 6.5 Temperature Adaption

Temperature is an additional, desired environmental parameter to adapt the ACS to. The presented concept of a deep learning model to adapt cross-sections to temperature is not advisable in this case. Especially for greater temperature ranges, additional information would be needed to adjust the cross-section successfully. This information corresponds to the partition function, which outlines changes in the population density of the two states participating in a transition. This effect has been explained previously with regard to Equation 12 and is again schematically visualized in Figure 43. Here the temperature $T_1$ is smaller than $T_2$, which results in a higher population density in the higher energy state $s_2$ compared to the lower energy state $s_1$. This change in population densities results in a change in line strengths and a different distribution in the resulting spectrum. In an extreme case, some transition is no longer possible as its ground state becomes depleted. This shows in the cross-section as the vanishing of a particular peak. Therefore, those probabilities and their changes differ for every transition and molecule and can not be generalized from a training set of different molecules. Thus, integrating changes over greater temperature ranges is impossible with the current architecture, as important information needed for this adaption is not included in the input or training data.

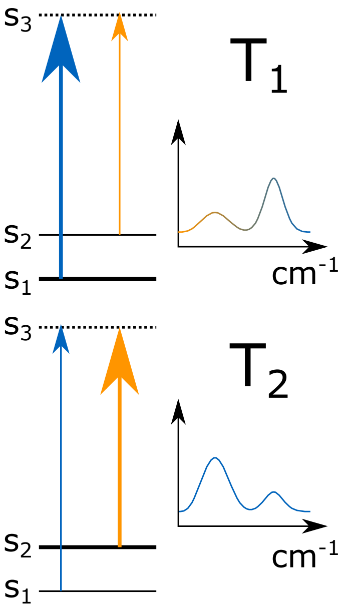


Figure 43: Schematic visualization of the effect of a temperature increase ($T_1 < T_2$) on the population density (left) and the resulting spectrum (right).

Any approach to enable the integration of temperature changes into the current model needs to integrate additional information into the model. In EPLL computation, this information is combined via the partition function, which is not available for every molecule in question. Knowledge about the molecule in question is required for the deep learning model to predict the spectral changes due to temperature successfully. The typical way to include knowledge about a molecule into deep learning architectures is by molecular fingerprints [176], [177]. A wide range of molecular fingerprints exist, focused on various molecular attributes. Two comparatively simple examples are the Morgan [169], and the Molecular Access System (MACCS) fingerprint [178]. Including those fingerprints into the SE block similar to the delta pressure value of the presented architecture did not suffice to enable the models to predict spectral adaption with temperature changes correctly. A more intricate inclusion of molecular knowledge into the model is needed for a successful temperature change prediction. The two selected fingerprints were not optimized with the molecule's rovibrational states in mind but covered a broad range of molecular attributes. While, in theory, a deep learning model to predict the changes in cross-sections by temperature seems possible, the correct fingerprint, which contains all necessary information, is needed. Molecular fingerprinting itself is a wide field of research, lately driven by graph neural networks [177]. The incorporation of those effects is an interesting question for future research.

This chapter presented a deep learning approach trained on simulated data to adapt measured absorption cross-sections to other environmental configurations. While the concept works well for pressure increases, pressure decreases, and temperature adaptions requires the inclusion of further information into the system, which is an exciting path for future research. The deep learning system can be trained for a set pressure change or to allow continuous pressure adaption. While the continuous adaption is less accurate, the use cases for both approaches are outlined and compared to the current state of the art. This allows for a broader range of molecules and environmental configurations for synthetic photoacoustic data generation, which will be explained in the next chapter.

# 7 Synthetic Data

Synthetic and measured absorption spectra can be used as a basis to generate synthetic photoacoustic spectra. For this purpose, integrating the photoacoustic measurement principle in the simulation is required to ensure correctness. In the following, a methodology to incorporate important measurement parameters with absorption spectra to generate synthetic photoacoustic spectra is presented. The approach is tested on two photoacoustic datasets, one amplitude modulated intended for breath analysis and the other wavelength modulated intended for methane measurements under varied humidity conditions. The application of the generated synthetic spectra for different machine learning techniques is also presented and compared to the training on measured spectra.

The methodology presented in this thesis is to use synthetic absorption profiles to overcome the data scarcity problem in the application of machine learning in gas sensing. Other methods to tackle data scarcity include linear combination, data augmentation, and large measurement campaigns. Linear combination refers to the generation of additional data by scaled summation of measured signals. While measurement campaigns require a lot of effort, data augmentation, and linear combination can quickly reach their limits with regard to generalization, as they only cover a small domain, similar to the measured base spectra. In some cases, those techniques are also not applicable. The approach of applying synthetic data to train machine learning models is presented here across different photoacoustic gas sensing applications and use cases. The synthetic data needs to be adapted to each sensing setup and can be created and perturbed as required to provide an extensive database with known deviations. From this data, models can be trained and created to solve challenges in the sensing applications. To measure a photoacoustic spectrum multiple measurements at different set points need to be performed. Therefore, the spectrum is acquired by multiple single-point measurements. This contrasts with other approaches like dual-comb spectroscopy, which acquires a full spectrum simultaneously. This results in an even more time-consuming data acquisition. Thus, the benefit of applying machine learning for data generation is two-fold: it enhances training data volume and representation and reduces measurement time.

The abstract concept of synthetic data generation for photoacoustic spectra is described in the following. This will be used to describe the synthetic data generation in the two examples of actual photoacoustic measurement setups in Section 7.1 and 7.2. Those sections also hold more details on each data generation process and examples of how the generated data can be used to improve machine learning applications compared to pure usage of measured data.

For the synthetic Photoacoustic Spectroscopy (PAS) data creation, details of the signal acquisition concept and specific measurement system must be considered. An overview of the synthetic data generation for photoacoustic setups is presented in Figure 44. It follows a simplified physical signal cascade that affects each spectrum. A partitioning into nine steps can be performed, which will be explained in greater detail in the following:

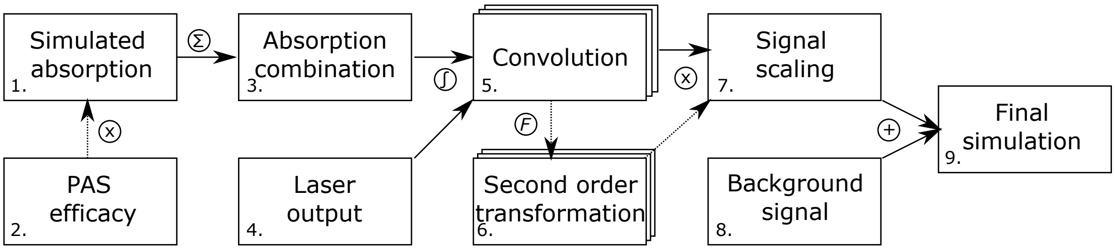


Figure 44: Visualization of the synthetic data generation for photoacoustic measurements.

1. **Generation of the absorption spectra of each molecule:** The absorption spectrum of each molecule can be computed from its ACS, which can be simulated from line databases or taken from measured reference data. To compute the absorption spectrum, additional variables like pressure, temperature, gas matrix, path length, and concentration have to be taken into account. The process of simulation from line databases is described in Section 4.1 and an adaption to environmental parameters for measured spectra in Section 6. This process yields one spectrum for each analyte in the gas mixture.

2. **Integration of the photoacoustic efficiency:** Since PAS does not measure the absorption directly but the pressure signal generated by the non-radiative relaxation of the molecule, only non-radiative relaxations contribute to this signal. When the bulk matrix is constant or has been found not to suffer from this non-spectral interference, this step can be skipped as the relaxation efficiency can be considered constant in this case. If that is not the case, the photoacoustic efficiency needs to be estimated via measurements or simulation as described in Section 3.4.4. Applying the efficiency factor over the full spectrum of a single molecule is a simplification, as the factor only corresponds to a single transition. This can be performed for simulated data but poses a problem when handling a measured ACS. Nevertheless, applying a single factor to the whole spectra performed well for the small spectral area under consideration in the experiments.

3. **Combination to a mixture spectrum:** Combining the molecule spectra into a mixture spectrum is performed by summation. This equals an omission of the mixture effects described by Hartmann et al. [31]. Nevertheless, this simplification is reasonable, as the mixture effects on the spectral shape are very minor and expensive to compute, especially considering the theoretical parameters, which are often unavailable and are not measurable with the sensing setup in question. This process now yields one mixture spectrum containing a combination of all analytes.

4. **Simulation of the laser output spectrum:** A precise simulation of the laser output spectrum is needed, as it greatly affects the final PAS spectrum. The parameterization of the laser driver can greatly impact the exact laser output spectrum. Mainly, when Amplitude Modulation (am)-PAS is performed, the 50 % duty cycle greatly changes the laser response from the Continous Wave (cw)-response that is typically provided by manufacturers. The simulation can be performed by fitting a typical laser output function to measured output spectra as described in the two examples in Subsections 7.1.1 and 7.2.1.

5. **Convolution of the laser output spectrum and the mixture spectrum:** The convolution of the laser output spectrum and the mixture absorption spectrum yields the signal generated at this laser set point. Since each signal acquisition combines many periods of signal generation through the lock-in-amplifier, an integration over all the wavenumbers can be performed. This yields one measurement point for a given set current, which can be directly used as the basis of a measurement point in Amplitude Modulated Photoacoustic Spectroscopy (am-PAS). As Wavelength Modulated Photoacoustic Spectroscopy (wm-PAS) is a second-order measurement, additional steps must be taken to simulate the second-order signal.

6. **Creation of the second order signal:** For the generation of a wm-PAS signal, the second-order information acquired by the setup needs to be simulated. To achieve a realistic simulation, the modulation of the laser around the center current is simulated in a fine grid, just as described in steps one to five. Those simulated data points of the modulation at each time step are combined, and a Fourier transformation is performed to extract the second-order information. The exact process is described in Subsection 7.2.1. This is repeated for each center current in the spectrum.

7. **Scaling of the simulated PAS spectrum:** Finally, a scaling of the simulated signal is performed. This scaling factor can be constant with the set current or contain additional terms as needed. This scaling factor combines multiple physical scaling effects that happen during signal acquisition. It covers not only microphone sensitivity and changes in optical power but also the resonance amplification of the cell (R-value) and the amplification by the Lock-in Amplifier (LIA). This scaling factor could also be adapted to depend on additional parameters. For example, the R-value of the cell can depend on the resonance frequency and speed of sound of the current bulk matrix. Therefore, this factor is most promising for improvement by integrating those factors.

8. **Simulation of the background signal:** The background signal and its simulation only apply to am-PAS and are governed mainly by two effects. The first one is the signal strength increase with laser current. Therefore, the optical power and the measured signal increase for each measurement point. The second effect is typically constant with respect to the optical power and originates in physical interactions of the beam at the cell windows and inside the cell. Both backgrounds can be easily obtained by performing a measurement without any analytes present.

9. **Generation of the final spectrum:** At last, to generate the final spectrum, the background and the scaled signal need to be added in the case of am-PAS. This simulated spectrum can now serve as the basis of model training.

Multiple parameters influence the synthetically generated signal during its generation, which correspond to real-world parameters, like the modulation depth in wm-PAS or the spectral width of the laser. Those parameters can be slightly adapted during simulation to correspond to real-world deviations during measurements. This way, a range of signals with known deviations can be created to be used during the training of machine learning models. The data generation described here can be parallelized to increase computation speed. In addition, all typical enhancements like data augmentation can be used. In the following two examples will be showcased: First, a photoacoustic setup for breath analysis that applies am-PAS using a Quantum Cascade Laser (QCL) and second, a wm-PAS setup involving an Interband Cascade Laser (ICL) which is aimed at methane detection for environmental analysis or natural gas monitoring.

## 7.1 Photoacoustic Setup for Breath Analysis

This section is partially based on our work published in [125], [179]–[181] and combines and extends the results presented there. The measurement setup is described after an introduction to breath analysis and the two analytes, acetone and ethanol. In Subsection 7.1.1, we combine the results from [180] and [125] to present a coherent data generation approach for this sensing setup which integrates system-specific parameter deviations. Next, an in-depth comparison of machine learning models trained on synthetic and measurement data, as well as a combination thereof, extended from [181] is presented in Subsection 7.1.2 and evaluated on synthetic and measured data in Subsections 7.1.3 and 7.1.4. The integration of synthetic data in variable selection is presented in Subsection 7.1.5.

Breath analysis as a diagnostic tool is considered a major opportunity in medicine. Non-invasive breath sampling is expected to reduce patient discomfort and thus enable higher patient compliance and more frequent health monitoring [4], [6], [182]. A broad range of analytical techniques are available for breath analysis, which have recently been thoroughly reviewed in [183] including a chapter focused on PAS [49]. PAS as a laser spectroscopic technique is considered a targeted, possibly real-time enabled analysis [19], [49], [184]. Targeted here refers to an a priori selection of the analytes of interest and a sensor design to most reliably detect those analytes.

In this work, the analytes of interest are acetone and ethanol. Acetone is a biomarker related to ketogenic diet [185]–[187] and acute decompensated heart failure [188]. An in-depth discussion of acetone as a biomarker can be found in [19, Chapter 2.1.4.]. The current study situation remains inconclusive to the quantitative, informative value of acetone as a biomarker due to large variations within the population [189]–[193]. Therefore, further large-scale studies are required. Concentrations in healthy individuals range from 0.3 to 0.9 ppmV [194]–[197].

Breath ethanol on the other hand can originate through external factors or the metabolism. It linearly correlates with blood ethanol concentration and can be used to assess driving ability. In conjunction with other biomarkers, it can be used to detect blood poisoning or liver disease [198]. Ethanol concentration in the breath of healthy humans ranges from 0.196-0.238 ppmV [194], [199]–[201].

Figure 45 visualizes the measurement setup concept. The input current is supplied by a Laser Driver Controller (LDC) 3736 from ILX Lightwave (Newport Corporation, US) to the QCL from AdTeach (AdTech Photonics Inc., US). The photoacoustic cell is a custom resonant design described in [53]. The acoustic signal is acquired through a Micro Electro Mechanical System (MEMS)-microphone ICS-40720 (InvenSens Inc., US) and amplified by a LIA Ametek 7270 (Ametek, US). The process is controlled using a custom LabVIEW program. The system is controlled to 35 °C to avoid condensation, but the laser is controlled to 23 °C for all measurements. The system, including the resonant photoacoustic cell and gas delivery system, are explained in detail in [53], [125], [179].

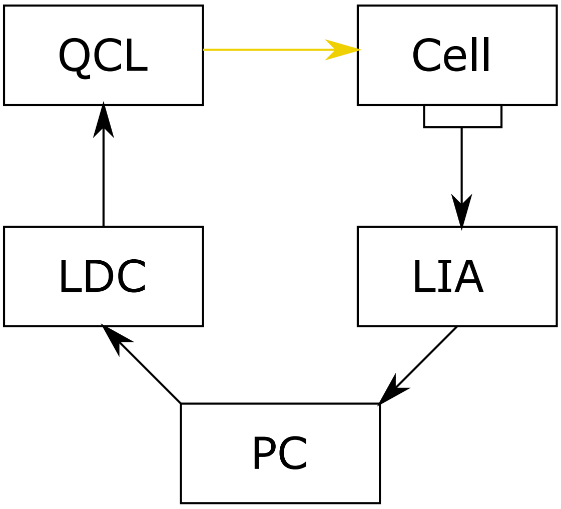


Figure 45: Setup of the QCL am-PAS measurement. A Personal Computer (PC) is used to control the Laser Driver Controller (LDC) which supplies current to the Quantum Cascade Laser (QCL). Light emitted by the QCL passes through the measurement cell. It is detected by a microphone and amplified by the Lock-in-Amplifier (LIA). the output is analyzed by the PC.

This setup used to generate the data within this section applies amplitude-modulated signal generation using a square wave signal with a 50 % duty cycle. The low–level of the square wave is kept at 250 mA, just below the laser threshold, so no light is produced during the off-phase. The laser is not entirely switched off to retain a more stable temperature, which conserves the laser and ensures a faster response. The high-level current of the square wave is modulated between 350 and 495 mA in steps of 1 mA to achieve different wavelengths. The QCL emits light in the range from 8263-8270 nm with a spectral resolution of 48 pm. This yields photoacoustic spectra with 146 measurement points each. The acquisition of one spectrum takes around 3 hours, highlighting the extreme data acquisition cost. The system reached an LoD of 0.25 ppbV for acetone and 6.47 ppbV for ethanol with an integration time of 10 s for a single point measurement [125].

Even though single-point measurements yield great sensitivity, they do not cope with spectral or non-spectral interferences in the photoacoustic measurement. This is especially of interest in a complex bulk matrix like human breath exhale. The full spectral information can be considered for each measurement to avoid incorrect measurements. Here, only a very small range of 7 nm is considered, similarly to [202], while others [118], [191], [203], [204] measured over a much broader range. Kistenev et al. [118], [204] applied Support Vector Machine (SVM) to predict diseases from the spectra directly but only had very limited samples (32 and less than 100). Therefore, the applicability of this study to medical practice is limited. We do not consider the problem of direct disease detection but want to predict the two analytes with high sensitivity and precision to enable future large-scale studies on the significance of those biomarkers.

The absorption spectra targeted by this sensing setup are presented in Figure 46. While a distinctive spectral pattern for water is visible, the other components don't show differentiated peaks in the area of interest. For $CO_2$, almost no absorption can be seen, as $CO_2$ is not a component of interest, and the spectral range was selected not to include $CO_2$ peaks. No distinctive structures can be found for acetone and ethanol, as only broad absorption bands can be found in this region. The primary analyte this setup was targeted to is acetone. Hence, a higher absorption coefficient for acetone can be seen in Figure 46. The simulated absorption spectra presented here are used as the basis for photoacoustic spectral simulation presented in the next section.

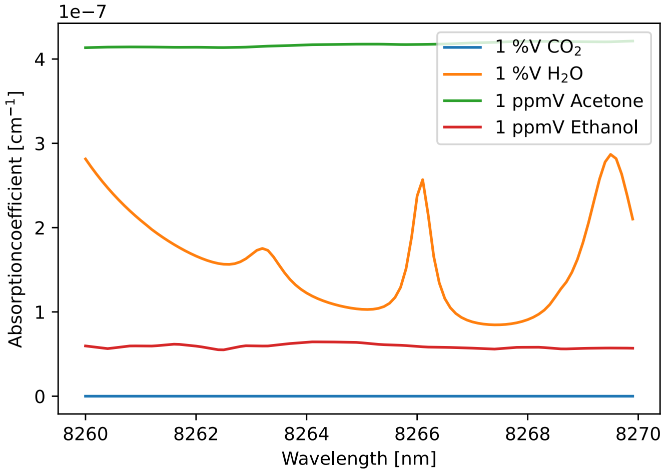


Figure 46: Absorption coefficient of the four components targeted with the am-PAS QCL setup for breath analysis. While $CO_2$ shows practically no absorption, distinctive structures can be found for $H_2O$. Acetone and ethanol show shallow absorption due to their broad structure.

### 7.1.1 Synthetic Data Creation

The synthetic data creation for the am-PAS QCL for breath analysis is intended to contain multiple physically relevant parameters to simulate deviations in the data stemming from, e.g., laser driver irregularities. We only simulate the signal amplitude, the phase is neglected. The process is described and evaluated in the following, and the parameters are presented.

To model the PAS spectrum of the am-PAS QCL, not all steps described in the main section are necessary. Step 2, the integration of the photoacoustic efficiency was omitted as previous measurements showed no change in photoacoustic efficiency for the expected bulk matrix. Step 6 was also omitted, as it only applies to wm-PAS.

To simulate the final signal, we follow an adapted PAS Equation 85. The measured photoacoustic signal $p_{\text{pas}}$ at each high-level current $I$ is computed, including the scaling factor $q(I)$ and background signal $bg(I)$. Uppercase letters with the subscript $\tilde{\nu}$ represent vectors over a range of wavelengths. The QCL output spectrum over all wavelengths $P_{\tilde{\nu}}(I)$ is convoluted with the absorption spectrum of the gas mixture $\sum_c A_{\tilde{\nu}c}$. The mixture absorption spectra are given by the sum over all components' absorption spectra $A_{\tilde{\nu}c}$.

$$p_{\text{pas}}(I) = q(I)\left[P_{\tilde{\nu}}(I) * \sum_c A_{\tilde{\nu}c}\right] + bg(I) \tag{85}$$

The single component absorption spectra of $CO_2$ and $H_2O$ were simulated using line data from HITRAN [14] and the Voigt equation [26] as implemented in HITRAN Application Programming Interface (hapi) [27]. Since acetone and ethanol are unavailable as line data, measured spectra from the Pacific Northwest National Laboratory (PNNL) library [28] were used. No additional adaption regarding temperature or pressure was required since the measured ACS were collected under similar circumstances. A conversion to absorption spectra was applied. An additional adaption of the ethanol base spectrum was performed as strong deviations concerning ethanol were noted. Since the ethanol gas tank only contained 96.5 % of the assigned concentration, ethanol spectra were multiplied with a factor of 0.965. This concentration error was validated by additional concentration measurements with an Fourier Transform Infrared (FTIR) (MKS Instruments, Inc; US). This error is still within the margins specified by the gas supplier (+/- 5 %) but led to noticeable errors in the synthetic data generation. In addition, the ethanol base spectrum was adapted between 8261-8264 nm. Finally, all spectra were scaled to the corresponding concentration and summed $\sum_c A_{\tilde{\nu}c}$.

The QCL output spectrum $P_{\tilde{\nu}}(I)$ for each high–level current proved to be relatively uncommon in this setup. It was computed using the handcrafted Equation 86. This instrument function consists of two components as described in Equation 86. The probability density function of a Gaussian $\mathcal{N}_{\tilde{\nu}}(\mu(I), b_0)$ and a handcrafted triangle function $Tri_{\tilde{\nu}}(I)$ as described in Equation 87. A laser output function is typically expected to form a Gaussian or Lorentzian curve centered at the output wavelength. The curve center depends on the high-level current supplied. In our case, it is computed using two adaptable parameters. The additional triangle function is required for this setup, as large deviations to the expected output function were observed. A non-neglectable signal at the lower energy side of the peak was observed, increasing with a higher high-level supply current as visualized in Figure 47. The actual laser output is most likely distorted in this way due to a limited slew rate of the laser driver. The adaptable parameters $b_{0-9}$ are fitted using measurements acquired with a 771 laser spectrum analyzer (Bristol Instruments Inc., US) for every supplied high–level current.

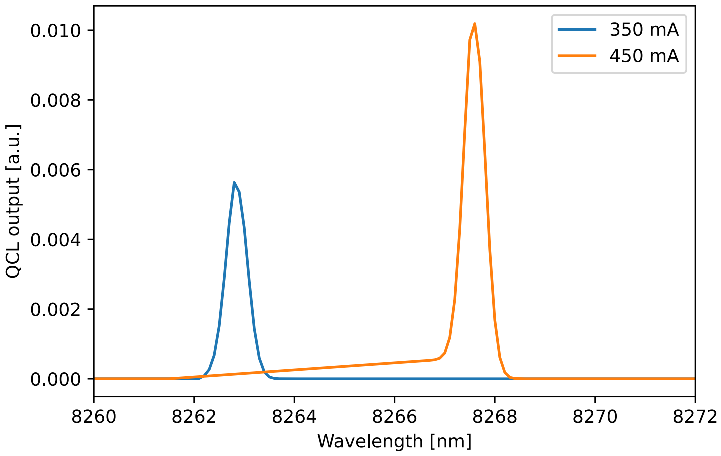


Figure 47: Simulated laser output at 350 and 450 mA high current. The 450 mA output shows the uncommon triangle-shaped signal.

$$P_{\tilde{\nu}}(I) = \mathcal{N}_{\tilde{\nu}}(\mu(I), b_0)(b_1 I + b_2) + Tri_{\tilde{\nu}}(I) \quad \text{with} \quad \mu(I) = b_3 I^2 + b_4 I + b_5 \tag{86}$$

$$Tri_{\tilde{\nu}}(I) = m/t \left(\vec{\tilde{\nu}} - (\mu - t)\right)_{\top 0}^{\perp m} \quad \text{with} \quad m = b_6 I + b_7 \quad \text{and} \quad t = b_8 I + b_9 \tag{87}$$

The laser output spectrum was convoluted with the mixture absorption spectrum. This step was repeated for all 146 measurement points. In the next step, the background signal was estimated from four independent PAS background measurements following Equation 88. The three parameters $c_{0-2}$ were kept constant after the fit.

$$bg(I) = c_0 I^2 + c_1 I + c_2 \tag{88}$$

Next, the signal transfer parameters $d_0$ and $d_1$ are estimated. Here only linear effects were assumed $q(I) = d_0 I + d_1$. Those parameters were fitted using eight spectra. Each spectrum was acquired with only one analyte. So two concentrations of $CO_2$, $H_2O$, acetone, and ethanol were used.

Finally, the QCL parameters $b_0$ and $b_5$ represent the standard deviation of the Gaussian function and the peak position needed adaption. This is due to the low resolution of our spectrum analyzer, which could not resolve the peak sufficiently. The same spectra used to fit $d_0$ and $d_1$ were used to optimize those two parameters.

Therefore, 15 adaptable parameters were fitted using a combination of twelve PAS spectra - two for each component and four background spectra - and measurements from the laser spectrum analyzer. Table 13 summarizes those parameters and their description. The parameters were further improved from their first publication in [180]. An additional twelve spectra that were not used to create the simulation were used to compute the validity of the simulation by computing the respective Mean Average Percentage

| Parameter | Value | Description | Estimated relative variance |
|---|---|---|---|
| $b_0$ | 0.21080 | HWHM of the QCL | - |
| $b_1$ | 2.4906e-05 | QCL output intensity linear with $I$ | - |
| $b_2$ | -5.8097e-03 | QCL output intensity constant | - |
| $b_3$ | 5.6738e-05 | instrument transfer quadratic with $I$ | - |
| $b_4$ | 2.0737e-03 | instrument transfer linear with $I$ | 0.27928* |
| $b_5$ | 8255.2 | instrument transfer constant | 7.0350e-06 |
| $b_6$ | 0.053663 | triangle slope linear with $I$ | - |
| $b_7$ | -18.0274 | triangle slope constant | - |
| $b_8$ | 2.5076e-06 | triangle height linear with $I$ | - |
| $b_9$ | -5.1060e-04 | triangle height constant | - |
| $c_0$ | -1.2022e-04 | background scaling quadratic with $I$ | - |
| $c_1$ | 0.24302 | background scaling linear with $I$ | 0.0048054 |
| $c_2$ | -55.228 | background scaling constant | 0.0097943 |
| $d_0$ | -3.6457e07 | PAS signal transfer scaling linear with $I$ | 0.024998 |
| $d_1$ | 3.2874e10 | PAS signal transfer constant | 0.012011 |

Table 13: Fitted parameters for the am-PAS QCL synthetic data generation, their values, estimated variance (where applicable), and physical correspondence. *estimation from the covariance matrix, which is unreliable due to strong correlations.

Error (MAPE) for each spectrum, which resulted in a mean MAPE score of 2.9 %. The results per spectrum and the respective concentrations are provided in Table 14.

Those spectra were also included, in addition to the spectra used to create the simulation, in an error contribution study. Here only the parameters $c_2$, $c_1$, $d_1$, $d_0$, $b_5$, and $b_4$ were analyzed. They were selected due to their strong correspondence with actual measurement parameters. The parameters $c_{1-2}$ define the linear and constant part of the background signal. The background signal can vary due to changes in the laser output power or slight misalignment of the setup. The parameters $d_{0-1}$ correspond to the linear and constant signal transfer or scaling. Those parameters can change mostly due to changes in the cell constant. Even though a check is performed before and after each spectral acquisition to ensure a measurement at the resonance frequency, the cell constant varies with changes in the bulk matrix as described in [125]. Finally, the parameters $b_{4-5}$ correspond to the linear and constant component of the instrument transfer function, e.g., the peak position at a certain high-level current $I$. Here, variations can be caused by the temperature compensation of the laser. Quadratic components were excluded from this error estimation.

The relative variance was computed from the covariance matrix and is provided in Table 13. This estimation disregards the covariance with other fitting parameters. Therefore, the feasibility of those

| **Measurement No.** | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **$H_2O$** [%V] | 0.89 | 1.24 | 0.89 | 1.24 | 0.89 | 1.24 |
| **$CO_2$** [%V] | 0 | 0 | 5 | 5 | 3 | 3 |
| **Ethanol** [ppmV] | 0 | 0 | 0 | 0 | 0 | 0 |
| **Acetone** [ppmV] | 3.92 | 3.92 | 3.92 | 3.92 | 1.96 | 1.96 |
| **MAPE score** [%] | 5.2 | 4.9 | 1.5 | 2.1 | 1.4 | 3.0 |

| Measurement No. | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|
| **$H_2O$** [%V] | 0 | 0.89 | 0 | 0.6 | 1.5 | 1.24 |
| **$CO_2$** [%V] | 0 | 3 | 0 | 0 | 1.5 | 5 |
| **Ethanol** [ppmV] | 7 | 7 | 4 | 6 | 2 | 4 |
| **Acetone** [ppmV] | 1.96 | 1.96 | 3.92 | 3.92 | 0.5 | 0 |
| **MAPE score** [%] | 2.6 | 1.3 | 2.7 | 5.5 | 2.2 | 2.0 |

Table 14: Concentrations of the measured evaluation spectra and their MAPE score with respect to the simulation.

estimates was confirmed by a Markov Chain Monte Carlo (MCMC) [205]. The MCMC showed a strong covariance of the parameter $b_4$, rendering the estimation from the covariance matrix less feasible and explaining the very high value of 18 %. For all other parameters, the simplified estimation of the relative variance was justified.

In Figure 48, the simulated PAS signal for pure water, acetone, and ethanol spectra is presented to provide an impression of the signal shape. While for water, two clear peaks are visible at 420 and 485 mA; no such clear signal is visible for acetone or ethanol. The upward slope in this non-preprocessed spectra results from the increased optical power for the increased high-level current. The main difference between the ethanol and acetone data is the difference in the slope. When the signal is compensated for the optical power change as presented in [125], a downward slope and a shallow elevation in ethanol can be seen. On the other hand, an upward slope of acetone is visible after compensation.

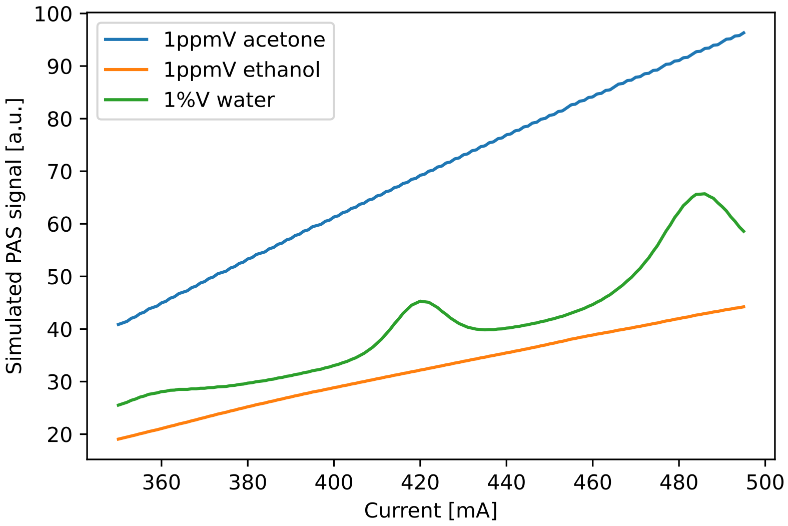


Figure 48: Simulated PAS spectrum for water, acetone, and ethanol. Adapted from [181].

For the generation of the training data, the standard deviation of the parameters $c_2$, $c_1$, $d_1$, $d_0$, $b_5$, and $b_4$ was scaled by $1/2$ and uniformly sampled during data generation. The spectra for a grid of concentrations were generated. Twenty samples each were equally spaced from 0-6 % water, 0-5 ppm acetone, and 0-10 ppm ethanol, and ten for 0-5 % carbon dioxide. This results in 80.000 synthetic spectra. Ten percent were randomly selected as the test set. The remaining samples were now used to train the machine learning models presented in the following sections.

### 7.1.2 Machine Learning Setup

In this section, we focus on a comparison between models trained on synthetic and measured data. Therefore, we employ different machine learning approaches previously presented in Section 5.2. To create a broad picture of the impact of the synthetic training data, we analyzed different machine learning algorithms representing different levels of complexity. Deep learning methods like CNNs [206] and recent transformer networks [100], which would constitute the next level of model complexity, were not included in the current study. All models used are summarized in Table 15 and were evaluated on the measured and the synthetic test set.

| Abbreviation | Model | Multi-/Uniresponse | Model complexity |
|---|---|---|---|
| MLR | Multi-Linear Regression | Multiresponse | Baseline |
| PLS1 | Partial Least Squares 1 | Uniresponse | Simple |
| PLS2 | Partial Least Squares 2 | Multiresponse | Simple |
| SVR | Support Vector Regression | Uniresponse | Advanced |
| ANN1 | Artificial Neural Network - 1 layer | Multiresponse | Advanced |
| ANN2 | Artificial Neural Network - 2 layer | Multiresponse | Advanced |
| RF | Random Forest | Uniresponse | Ensemble |
| GB | Gradient Boosting | Uniresponse | Ensemble |

Table 15: Overview of machine learning algorithms evaluated during synthetic data experiments.

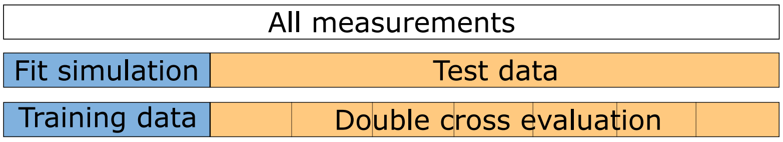


Figure 49: Indicated data partition for the measured am-PAS QCL data for the models trained on synthetic (top) and measured (bottom) data. The data used to fit the synthetic data model was never used to evaluate the measurement model to ensure comparability between the two approaches. Only the outer cross-validation splits are visualized.

The synthetic dataset, including simulated measurement deviations, was created as described in Subsection 7.1.1. The measurement dataset consists of 30 spectra, of which 22 were used in a seven-fold double cross-validation. The eight samples already used to fit the synthetic spectra were always attributed to the training data. This concept is visualized in Figure 49 and ensures a comparable evaluation for the models trained on the synthetic and measured datasets. The outer double cross-validation splits were created not to contain the same mixture of components in the training and test split. The inner cross-validation was fourfold. The concentrations are provided in Table 16, indicating the outer double cross-validation split. A separate hyperparameter tuning, including all measurement data, was performed to evaluate the synthetic data, again employing four-fold cross-validation.

Data augmentation is typically used when working with measured data to increase variation and, therefore, the robustness of the final model. We followed the approaches presented by Bjerrum et al. [141] and changed the slope and offset of the data, added random noise and a small multiplicative factor. In contrast to other works [12], [141] we did not derive the intensity of our data augmentation from the dataset. We expected the best amount of data augmentation to differ by model and, therefore, included the data augmentation parameters in the Bayesian hyperparameter tuning. All models' corresponding tuning ranges were kept constant and are provided in Table 17. The possible tuning ranges for the model hyperparameters can be found in Tables 18 to 22. For an explanation of the hyperparameters used, the reader is referred to the documentation of the scikit learn package (version 1.2.2) [147]. All models were

| **Sample** | **Acetone** [ppmV] | **Ethanol** [ppmV] | **Water** [%V] | $\mathbf{CO_2}$ [%V] | **Split** |
|---|---|---|---|---|---|
| 1 | 3.92 | 0 | 0.89 | 0 | 0 |
| 2 | 3.92 | 0 | 1.24 | 0 | 1 |
| 3 | 0 | 0 | 0.89 | 0 | / |
| 4 | 0 | 0 | 1.24 | 0 | / |
| 5 | 3.92 | 0 | 0 | 5 | 4 |
| 6 | 0 | 0 | 0 | 5 | / |
| 7 | 3.92 | 0 | 0 | 0 | / |
| 8 | 3.92 | 0 | 0 | 5 | 4 |
| 9 | 3.92 | 0 | 0.89 | 5 | 0 |
| 10 | 3.92 | 0 | 1.24 | 5 | 1 |
| 11 | 3.92 | 0 | 0 | 5 | 4 |
| 12 | 3.92 | 0 | 0.89 | 5 | 0 |
| 13 | 3.92 | 0 | 1.24 | 5 | 1 |
| 14 | 1.96 | 0 | 0 | 0 | / |
| 15 | 1.96 | 0 | 0 | 3 | 5 |
| 16 | 1.96 | 0 | 0.89 | 3 | 0 |
| 17 | 1.96 | 0 | 1.24 | 3 | 6 |
| 18 | 0 | 0 | 0 | 3 | / |
| 19 | 0 | 0 | 0.89 | 3 | 2 |
| 20 | 1.96 | 0 | 1.24 | 3 | 6 |
| 21 | 0 | 0 | 0.89 | 5 | 2 |
| 22 | 0 | 0 | 1.24 | 5 | 3 |
| 23 | 0 | 7 | 0 | 0 | / |
| 24 | 1.96 | 7 | 0 | 0 | 3 |
| 25 | 1.96 | 7 | 0.89 | 3 | 5 |
| 26 | 0 | 4 | 0 | 0 | / |
| 27 | 3.92 | 4 | 0 | 0 | 2 |
| 28 | 3.92 | 6 | 0.6 | 0 | 3 |
| 29 | 0.5 | 2 | 1.5 | 4 | 5 |
| 30 | 0 | 4 | 1.24 | 5 | 6 |

Table 16: Concentrations of the measurement data set and their outer double cross-validation split.

subject to Bayesian hyperparameter tuning as implemented by Weights and Biases [146] for 100 epochs. Here, it has to be noted that the models trained on the synthetic dataset were tuned with respect to the synthetic validation data. In contrast, the measurement and augmented measurement models were tuned towards their cross-validation sets.

| Parameter | Min | Max | Distribution |
|---|---|---|---|
| Multiply | e-15 | e-2 | Log uniform |
| Noise | e-20 | e-2 | Log uniform |
| Offset | e-15 | e-2 | Log uniform |
| Sample num | 1 | 166 | Int uniform |
| Slope | e-15 | e-2 | Log uniform |

Table 17: Hyperparameter ranges for data augmentation.

| Parameter | Min | Max | Categories | Distribution |
|---|---|---|---|---|
| n components | 3 | 100 |  | Int uniform |

Table 18: Hyperparameter ranges for PLS1 and PLS2.

| Parameter | Min | Max | Categories | Distribution |
|---|---|---|---|---|
| C | -15 | 1 |  | Log uniform |
| Coef0 | 0 | 2 |  | Uniform |
| Degree | 1 | 5 |  | Int uniform |
| Epsilon | 0.01 | 0.6 |  | Uniform |
| Gamma |  |  | Auto, Scale | Categorical |
| Kernel |  |  | Linear, Rbf, Poly, Sigmoid | Categorical |
| Shrinking |  |  | True, False | Categorical |

Table 19: Hyperparameter ranges for SVR.

| Parameter | Min | Max | Categories | Distribution |
|---|---|---|---|---|
| Activation |  |  | Identity, Logistic, Tanh, Relu | Categorical |
| Alpha | e-15 | e-2 |  | Log uniform |
| Early stopping |  |  | True, False | Categorical |
| Hidden layer1 | 5 | 500 |  | Int uniform |
| Hidden layer2 | 5 | 500 |  | Int uniform |
| Learning rate |  |  | Constant, Invscaling, Adaptive | Categorical |
| Learning rate init | e-15 | e-1 |  | Log uniform |
| Solver |  |  | LBFGS, SGD, Adam | Categorical |

Table 20: Hyperparameter ranges for ANN1 and ANN2.

| **Parameter** | **Min** | **Max** | **Categories** | **Distribution** |
|---|---|---|---|---|
| Max depth | 2 | 20 | | Int uniform |
| Max features | | | Auto, Sqrt, Log2 | Categorical |
| Min samples leaf | 1 | 10 | | Int uniform |
| Min samples split | 2 | 20 | | Int uniform |
| N estimators | 10 | 200 | | Int uniform |

Table 21: Hyperparameter ranges for RF.

| **Parameter** | **Min** | **Max** | **Categories** | **Distribution** |
|---|---|---|---|---|
| Criterion | | | Sqared error, Friedman mse, Absolute error | Categorical |
| Learning rate | e-15 | e-1 | | Log uniform |
| Loss | | | Squared error, Absolute error, Huber, Quantile | Categorical |
| Max depth | 2 | 30 | | Int uniform |
| Max features | | | Auto, Sqrt, Log2 | Categorical |
| Min samples leaf | 1 | 10 | | Int uniform |
| Min samples split | 2 | 20 | | Int uniform |
| N estimators | 3 | 200 | | Int uniform |
| Subsample | 0.85 | 1 | | Uniform |

Table 22: Hyperparameter ranges for GB.

### 7.1.3 Machine Learning Evaluation with Synthetic Data

The evaluation results on the synthetic dataset are visualized as a heatmap in Figure 50. The different machine learning models are listed on the vertical axis, while the different training datasets are presented on the horizontal axis. The model performance is presented in the grid and color coded. In this case, it is computed on the synthetic test dataset. It is presented as the scaled sum-Mean Absolute Error (MAE). This performance score was chosen to represent one metric even though the concentrations of the analytes vary greatly. For this score, each component's MAE was computed, scaled by the maximal concentration of this component: 6 % for water, 5 ppm for acetone, and 10 ppm for ethanol, and subsequently summed. A lower score indicates better performance. Analyzing Figure 50, a devastating performance of the models trained on the measurement or augmented measurement dataset in the middle and rightmost column can be seen. They show a scaled sum–MAE between 36 and 91 % for every model, while the models trained on the synthetic dataset only show a scaled sum–MAE of 2.5 to 3 % for all models except the ensemble models. Those perform worse than expected with 13 and 20 % scaled sum–MAE. The worse performance of the measurement and augmented models indicates that the synthetic dataset contains additional perturbations that are not provided in the measurement dataset.

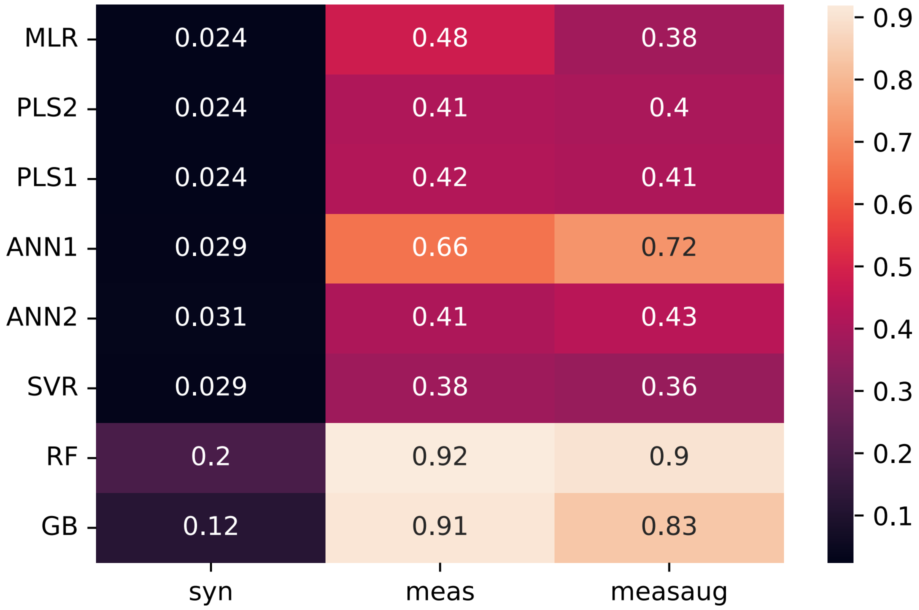


Figure 50: Scaled sum-MAE [a.u.] on the synthetic test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

The same plot is provided per molecule in Figures 51 to 53 for a more in-depth evaluation. Here the scaled MAE for each molecule is used to provide comparability among the analytes. The sum of all moleclues' scaled MAE corresponds to the scaled sum-MAE. In Tables 23 to 25, the MAE (unscaled), as well as the standard deviation of the models for water, acetone, and ethanol, is provided. For water in Table 23, we can see a MAE of less than 1 % for most models trained on the measurement and augmented dataset on this molecule, which indicates that the position and shape of the prominent water features are well captured on the synthetic dataset. The bad performance of the ensemble models with a MAE of more than 2.4 %V has to be noted. The ensemble methods did not improve the model's capability to transfer to this different domain, which is visible for all three components from far worse performances than other models.

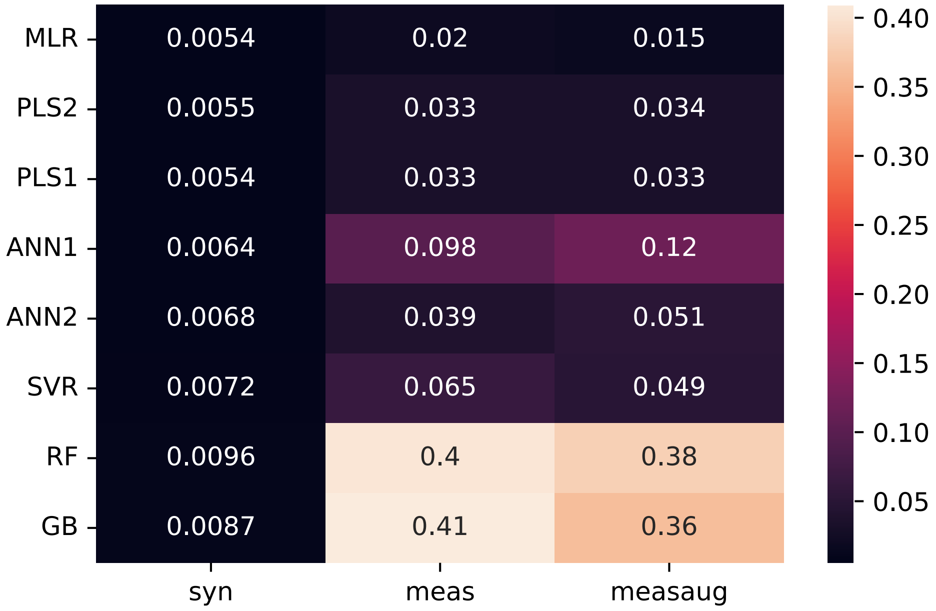


Figure 51: Scaled MAE [a.u.] of water on the synthetic test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

When focusing the evaluation on acetone, we again see the synthetic training performing better as was expected. But different to the evaluation on water, a strong difference between the measurement and augmented dataset can be seen in Figure 52 and Table 24 with the augmented dataset performing a lot worse when evaluated on synthetic data. While for the measurement dataset, the MAE is generally below 1 ppmV, it is above 2 ppmV for all augmented models. This might be attributed to a part of the augmentation which hinders the correct acetone prediction. Since acetone shows a relatively shallow slope without visible features in the spectrum, this might be disturbed by the slope augmentation procedure.

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 0.0324 ± 0.0270 | 0.118 ± 0.0988 | 0.0916 ± 0.0864 |
| PLS2 | 0.0327 ± 0.0273 | 0.201 ± 0.215 | 0.201 ± 0.215 |
| PLS1 | 0.0324 ± 0.0270 | 0.198 ± 0.213 | 0.200 ± 0.214 |
| ANN1 | 0.0385 ± 0.0311 | 0.590 ± 0.659 | 0.707 ± 0.763 |
| ANN2 | 0.0407 ± 0.0341 | 0.233 ± 0.292 | 0.304 ± 0.302 |
| SVR | 0.0435 ± 0.0327 | 0.388 ± 0.306 | 0.292 ± 0.295 |
| RF | 0.0574 ± 0.0524 | 2.42 ± 1.63 | 2.28 ± 1.62 |
| GB | 0.0708 ± 0.0631 | 2.46 ± 1.65 | 2.17 ± 1.63 |

Table 23: MAE on water and its standard deviation in %V on the synthetic test dataset for all training datasets and models.

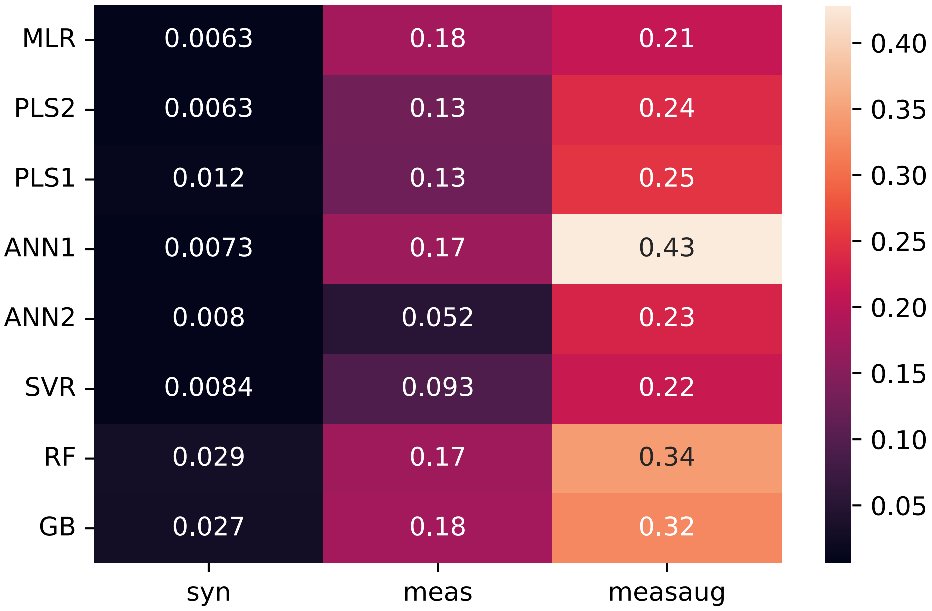


Figure 52: Scaled MAE [a.u.] of acetone on the synthetic test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

Finally, the ethanol evaluation on the synthetic dataset shows the reverse effect of augmentation compared to acetone. This is depicted in Figure 53 and Table 25. While again, the models trained on the measurement and augmented measurement dataset perform far worse than the synthetically trained models, here, the augmented dataset shows far better performance for ethanol prediction compared to the models trained on the measurement dataset. For ethanol, the models trained on the measurement dataset show a MAE above 2 ppmV, while the models trained on the augmented dataset show a MAE below 1 ppmV. Here, the augmentation has greatly improved the performance of the models. Also, the

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 0.0313 ± 0.0258 | 0.895 ± 0.750 | 2.09 ± 1.78 |
| PLS2 | 0.0314 ± 0.0258 | 0.647 ± 0.546 | 2.39 ± 1.98 |
| PLS1 | 0.125 ± 0.107 | 0.633 ± 0.527 | 2.49 ± 2.07 |
| ANN1 | 0.0364 ± 0.0308 | 0.855 ± 1.07 | 4.29 ± 4.70 |
| ANN2 | 0.0399 ± 0.0340 | 0.262 ± 0.199 | 2.31 ± 2.09 |
| SVR | 0.0420 ± 0.0336 | 0.463 ± 0.366 | 2.15 ± 1.74 |
| RF | 0.144 ± 0.113 | 0.867 ± 0.592 | 3.44 ± 2.52 |
| GB | 0.146 ± 0.115 | 0.888 ± 0.623 | 3.24 ± 2.37 |

Table 24: MAE on acetone and its standard deviation [ppmV] on the synthetic test dataset for all training datasets and models.

bad performance of the ensemble methods trained on the synthetic dataset has to be noted for ethanol prediction. While all other models show a MAE below 0.2 ppmV, the Gradient Boosting (GB) has a MAE of 0.876 ppmV, and the Random Forest (RF) even of 1.58 ppmV, which indicates a failure of the ensemble models to capture the actual influence of ethanol on the spectra.

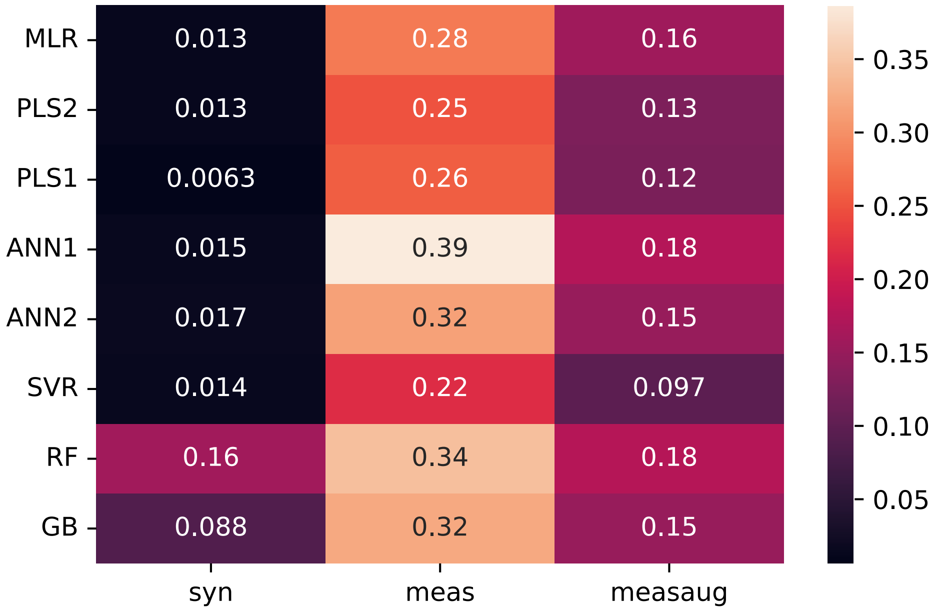


Figure 53: Scaled MAE [a.u.] of ethanol on the synthetic test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 0.125 ± 0.107 | 2.80 ± 2.27 | 0.785 ± 0.667 |
| PLS2 | 0.125 ± 0.107 | 2.49 ± 2.06 | 0.634 ± 0.533 |
| PLS1 | 0.0313 ± 0.0258 | 2.58 ± 2.14 | 0.619 ± 0.512 |
| ANN1 | 0.150 ± 0.130 | 3.87 ± 4.00 | 0.876 ± 1.08 |
| ANN2 | 0.165 ± 0.140 | 3.15 ± 2.72 | 0.746 ± 0.635 |
| SVR | 0.138 ± 0.115 | 2.18 ± 1.65 | 0.486 ± 0.386 |
| RF | 1.58 ± 1.30 | 3.44 ± 2.43 | 0.882 ± 0.614 |
| GB | 0.876 ± 0.639 | 3.21 ± 2.35 | 0.746 ± 0.528 |

Table 25: MAE on ethanol and its standard deviation [ppmV] on the synthetic test dataset for all training datasets and models.

The evaluation of all three models on the synthetic dataset allows to draw the following conclusions:

- All models trained on the synthetic dataset perform better than those trained on measurement and augmented data when evaluated on the synthetic test data. This is expected as they were trained on data from the same distribution.
- The simulation of the water spectra is very close to the measured spectra as the error from the models trained on the measurement and augmented dataset are comparatively small for the evaluation on synthetic data.
- The ensemble models trained on synthetic data fail to predict ethanol concentrations even on the same dataset correctly.
- The models trained on the augmented dataset perform worse than the models trained on the measurement dataset for acetone prediction.
- In contrast, training on augmented data improves the prediction of ethanol in comparison to non-augmented measurement data.
- While the prominent features of water seem to be captured well in the synthetic dataset, the subtle slopes and features of ethanol and acetone are not as well captured, leading to worse performances of the models trained on measured and augmented data for those molecules.
- The standard deviation of all models and all training datasets is high, showing a correlation to the MAE.

### 7.1.4 Machine Learning Evaluation with Measurement Data

A similar evaluation was chosen for the measurement test set. In contrast to the synthetic test set, the evaluation on the measurement data was a double cross-validation. The data was split into seven outer splits as indicated in Figure 49 and double cross-validated with four inner folds for 100 epochs with Bayesian hyperparameter tuning. An overview of the scaled sum-MAE scores of each model and training dataset evaluated on the measurement test data is presented in Figure 54. The different machine learning models are listed on the vertical axis, while the different training datasets are presented on the horizontal axis. The model performance is presented in the grid and color coded. In this case, it is computed over all outer measurement dataset splits or the measurement test set for the models trained on the synthetic dataset. It is presented as the scaled sum-MAE. Here, a reverse trend as on the synthetic data evaluation shown in Figure 50 can be seen, with the measurement and augmented measurement training dataset performing better than the synthetic training dataset. Such performance can be attributed to the domain difference between the synthetic and measured or augmented measurement dataset. The ensemble methods, RF and GB, trained on the measurement dataset, performed worse than the other models, similar to the synthetically trained ensemble models on the synthetic dataset.

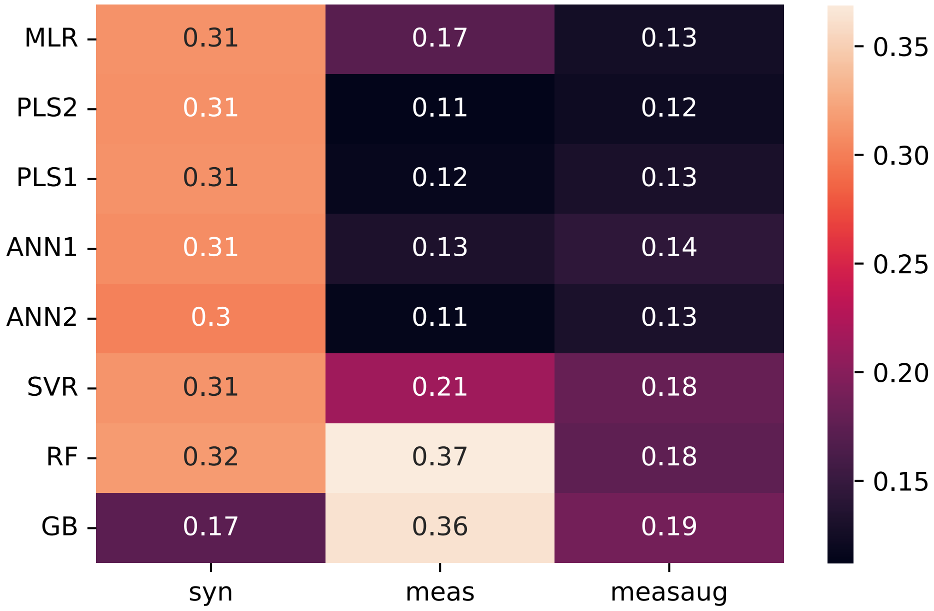


Figure 54: Scaled sum-MAE [a.u.] on the measurement test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

The evaluation of the prediction of the water concentration is presented in Figure 51 with respect to the scaled MAE and in Table 26 the MAE in %V as well as the standard deviation is displayed. Here, the worse performance of the ensemble methods, RF and GB, trained on the measurement and measurement augmented dataset becomes visible. All models except those show a MAE below 0.2 %V, with the synthetically trained models showing around double the error than their respective counterparts trained on the measurement or augmented measurement dataset.

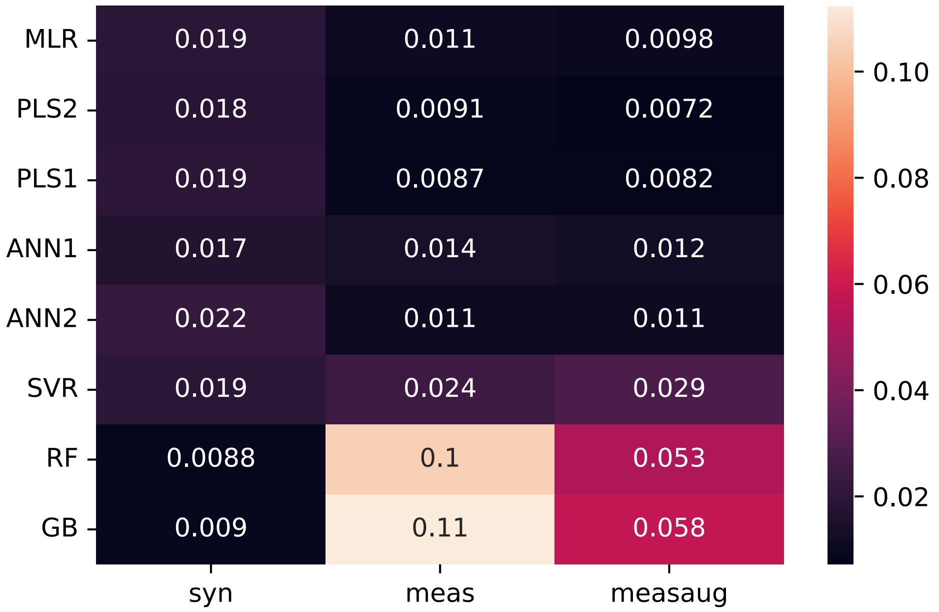


Figure 55: Scaled MAE [a.u.] of water on the measurement test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 0.114 ± 0.164 | 0.0649 ± 0.0688 | 0.0586 ± 0.0738 |
| PLS2 | 0.111 ± 0.152 | 0.0547 ± 0.0356 | 0.0432 ± 0.0388 |
| PLS1 | 0.115 ± 0.163 | 0.0524 ± 0.0371 | 0.0492 ± 0.0545 |
| ANN1 | 0.102 ± 0.129 | 0.0832 ± 0.0782 | 0.0742 ± 0.0888 |
| ANN2 | 0.130 ± 0.150 | 0.0646 ± 0.0501 | 0.0646 ± 0.0778 |
| SVR | 0.114 ± 0.124 | 0.171 ± 0.200 | 0.171 ± 0.122 |
| RF | 0.0531 ± 0.0591 | 0.628 ± 0.284 | 0.318 ± 0.418 |
| GB | 0.0785 ± 0.0748 | 0.674 ± 0.273 | 0.318 ± 0.438 |

Table 26: MAE on water and its standard deviation in %V on the measurement test dataset for all training datasets and models.

The acetone prediction presents an advantage of some simple and complex models (Multi Linear Regression (MLR), Partial Least Squares (PLS) 1&2, ANN 1&2) trained on the measured dataset with a MAE below 0.2 ppmV as can be seen in Table 27 and from Figure 56. The ensemble methods trained on the synthetic dataset show similarly good performance, while all other models trained on the synthetic dataset show a MAE of around 0.35-0.4 ppmV. The models trained on the augmented dataset show an even worse performance with a MAE from 0.77 to 1.09 ppmV. A similar trend could already be seen during the evaluation on the synthetic test dataset, where the models trained on the augmented measurement dataset also showed a worse performance for acetone.

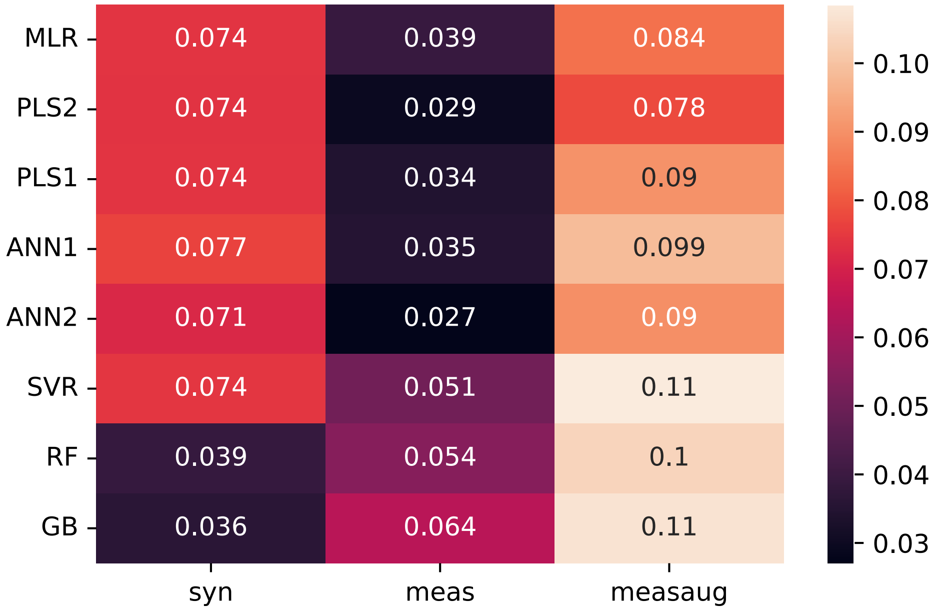


Figure 56: Scaled MAE [a.u.] of acetone on the measurement test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

Finally, the ethanol evaluation on the measurement test dataset is depicted in Figure 57 and the MAE in ppmV and its standard deviation are provided in Table 28. The main contribution to the higher scaled sum-MAE of the synthetic models seen in figure 54 can be attributed to the ethanol prediction with a MAE above 2 ppmV for all models except GB. In addition, a similar trend as in the evaluation on the synthetic test dataset can be seen, as the models trained on the augmented dataset perform better with a MAE below 0.2 ppmV for most models compared to a MAE above 0.7 ppmV of the models trained on the measurement dataset.

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 0.371 ± 0.238 | 0.194 ± 0.229 | 0.843 ± 1.17 |
| PLS2 | 0.369 ± 0.237 | 0.146 ± 0.121 | 0.778 ± 0.876 |
| PLS1 | 0.371 ± 0.238 | 0.171 ± 0.118 | 0.901 ± 1.15 |
| ANN1 | 0.383 ± 0.277 | 0.176 ± 0.124 | 0.985 ± 1.10 |
| ANN2 | 0.357 ± 0.276 | 0.135 ± 0.0918 | 0.896 ± 1.05 |
| SVR | 0.371 ± 0.251 | 0.253 ± 0.237 | 1.09 ± 1.43 |
| RF | 0.193 ± 0.154 | 0.272 ± 0.468 | 1.04 ± 1.65 |
| GB | 0.178 ± 0.138 | 0.320 ± 0.611 | 1.07 ± 1.97 |

Table 27: MAE on acetone and its standard deviation [ppmV] on the measurement test dataset for all training datasets and models.

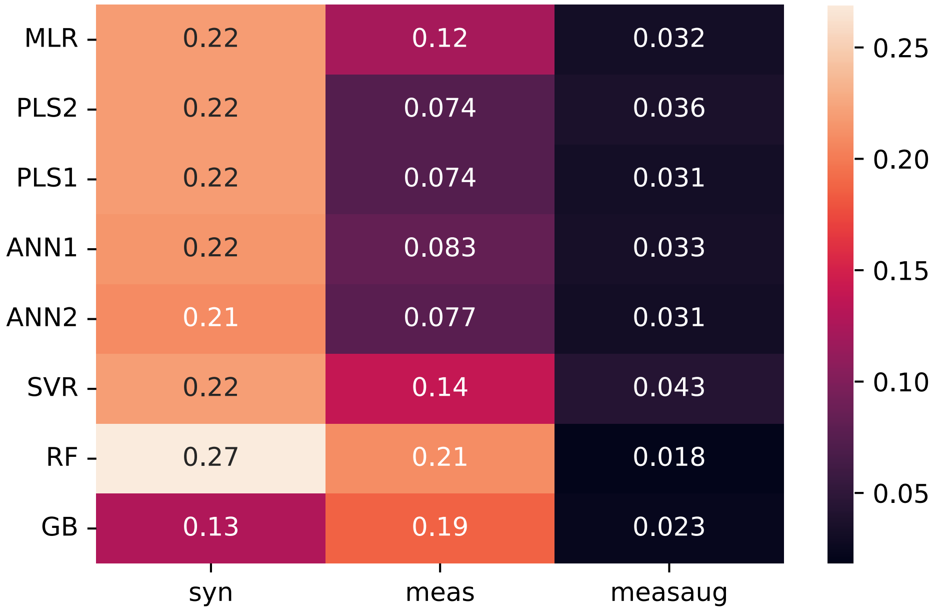


Figure 57: Scaled MAE [a.u.] of ethanol on the measurement test dataset per model and training dataset (synthetic, measured, measured-augmented). Lower values indicate better performance.

| Model | Synthetic | Measurement | Augmentation |
|---|---|---|---|
| MLR | 2.18 ± 1.17 | 1.22 ± 1.38 | 0.159 ± 0.199 |
| PLS2 | 2.18 ± 1.16 | 0.738 ± 0.734 | 0.181 ± 0.143 |
| PLS1 | 2.19 ± 1.17 | 0.738 ± 0.719 | 0.157 ± 0.150 |
| ANN1 | 2.15 ± 1.22 | 0.825 ± 0.845 | 0.166 ± 0.188 |
| ANN2 | 2.09 ± 1.25 | 0.771 ± 0.670 | 0.154 ± 0.184 |
| SVR | 2.20 ± 1.15 | 1.39 ± 1.52 | 0.215 ± 0.167 |
| RF | 2.69 ± 1.94 | 2.10 ± 1.71 | 0.0919 ± 0.342 |
| GB | 1.26 ± 0.754 | 1.87 ± 2.43 | 0.116 ± 0.386 |

Table 28: MAE on ethanol and its standard deviation [ppmV] on the measurement test dataset for all training datasets and models.

The conclusions to be drawn from this evaluation on the measurement test set can be summarized as follows:

- The models trained on the measurement and augmented measurement dataset perform better than the synthetically trained models as the domain gap persists, and the synthetic models don't generalize well to the new domain.
- Similar to the evaluation on the synthetic dataset, an improvement in ethanol prediction after data augmentation can be seen.
- The prediction of acetone seems to be hindered by the data augmentation as models perform worse when trained on the augmented dataset than the measurement dataset.
- The ensemble models trained on the measurement dataset perform worse on all three components compared to the other models trained on the same dataset.
- The standard deviation of all models and all training datasets is high, showing a similar value to the MAE.
- The ensemble methods trained on the synthetic dataset show a competitive performance to the models trained on the measurement dataset on acetone and water prediction.

Overall, four main conclusions concerning this dataset can be drawn:

1. **Simple machine learning methods suffice.**
   The most important conclusion from the evaluation presented is that already simple machine learning methods like PLS suffice to predict concentrations from the presented data. While on the synthetic dataset, even the MLR as a baseline method can reach one of the best overall results, on the measurement dataset, this baseline MLR is outperformed by simple machine learning methods like PLS, each trained on the respective training dataset. The need for a more intricate method on the measurement dataset can be attributed to the higher deviations between the data due to measurement effects or to overfitting of the MLR. The overall linearity of the dataset, so no changes to the relaxation efficiency of the molecules in the targeted matrix, has already been shown

in [53]. Therefore, no strong non-linear effects show in the presented dataset, and hence, no complex non-linear models like ANNs are needed to successfully predict the concentrations from the dataset.

With a more complex dataset, the effect of overfitting could become a more pressing problem, which a large amount of synthetic data can alleviate. Still, for a simple dataset as this, a PLS trained on the measured dataset performs best.

2. **The synthetic data does not yet fully cover the measurement domain.**
   The strong differences in performance of the models trained on the different domains show the strong domain gap between the synthetic and measured dataset. The models trained on the synthetic dataset perform worse when evaluated on the measurement dataset and vice versa. A further improvement of the synthetic data generation could improve this problem. The primary possibility for improvement is seen in the underlying ethanol absorption spectrum, which shows a comparatively low resolution - regarding wavelengths and signal height. This has already been discussed in depth in our previous work [125].

   In addition, domain adaption methods can be used to leverage the large amount of synthetic data and the measured dataset. This includes, for example, the fine-tuning of ANNs as was presented on this dataset in [181] or a hybrid training approach where the synthetic and measurement samples are combined during the adaption process of the model.

3. **Data augmentation on the small measurement dataset has to be used with care.**
   While data augmentation shows excellent improvements compared to the pure measurement dataset for ethanol prediction, a worse result on acetone prediction was visible for both test datasets. Therefore, when using data augmentation in spectral regression, an in-depth evaluation is needed, and in this case, both components' predictive quality has to be kept in mind.

4. **High standard deviations hint towards possible improvements in the measurement setup.**
   For the synthetic and the measurement models, high standard deviations in the range of the MAE can be seen for all components and models. Since this effect also shows on the synthetic dataset, which employs the modeled measurement setup deviations, this high deviation indicates room for improvement in the measurement setup. When measurements are more reproducible and the deviation, for example, of the light source, can be reduced, a lower standard deviation can be reached.

In this chapter, the synthetic am-PAS QCL spectra were evaluated as training datasets for different machine learning models compared to a small measured and augmented measurement dataset. The evaluation showed that simple machine learning models like PLS suffice to predict water, acetone, and ethanol concentration from the measured spectrum as no strong non-linearities are included in this dataset. The domain gap between the synthetic and measurement dataset is still too large for the models trained on the synthetic dataset to be used directly on measured spectra. In the following chapter, synthetic spectra are employed for variable reduction to shorten the measurement time of the sensing setup.

### 7.1.5 Variable Reduction with Synthetic Data

Data scarcity becomes even more pronounced if the spectrum in question contains fewer measurement points as the risk of overfitting increases. Since in am-PAS, each measurement point corresponds to an increase in measurement time, a reduction is desirable. Each spectral measurement presented in the previous section took around three hours of measurement time. Hence, a selection of the most valuable variables to maintain predictive quality while reducing the measurement time was sought out. In this section, the influence of the underlying dataset is evaluated. We select the variables either from the synthetic or the measured dataset and also train the resulting models on both kinds of datasets. Hence, conclusions can be drawn on the influence of synthetic data in the case of variable reduction.

Since the PAS measurements have a very high resolution of 7 nm, we expected a high collinearity of the data. In addition, three components needed to be predicted from the selected variables. As such, we chose the variable selection method Covariance Selection (CovSel) [153], which is known to handle high collinearity and can be used with multi-response data. Section 5.6 provides an overview of variable selection algorithms. While there are more methods to handle the requirements of this task, a comparison of different variable selection algorithms was beyond the scope of this work.

The results of CovSel applied to the measured and synthetic dataset are presented in Figure 58 and Table 29. The plot depicts the first five selected variables for each dataset. Their order is indicated at the bottom. For a more straightforward interpretation in the background, the simulated spectra for 1 ppmV of acetone and ethanol and 1 %V water are plotted. The shaded area around the spectra shows the normalized selectivity ratio [207]. Here, a larger shaded area corresponds to a higher variable importance. The high selectivity for ethanol from 370 - 410 mA stems from a shallow peak, which only becomes evident after post-processing the data. Even though some selected positions are very close or even the same for both datasets (1 measured & 2 synthetic, 3 measured & 5 synthetic), overall, a different selection of positions becomes evident. Nevertheless, subsequent experiments showed a minor influence of the selected positions on the overall performance. This is probably due to ethanol and acetone's broad spectral structures, which allow similarly good quantification at multiple positions.

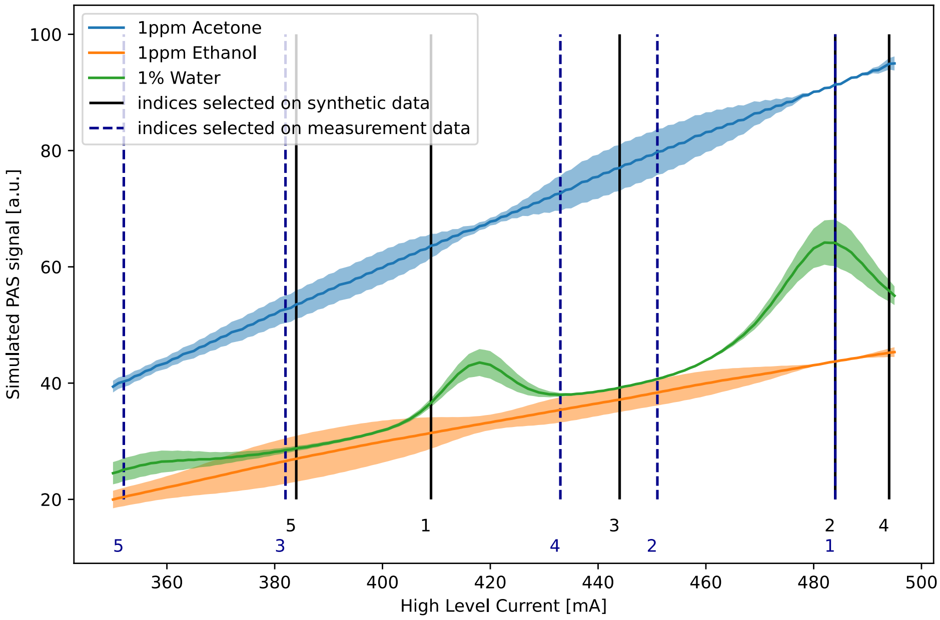


Figure 58: The five most informative spectral positions selected on synthetic and measurement data. In the background simulated example spectra of acetone, ethanol, and water with their indicated importance values are depicted. A larger shaded area indicates high importance at a spectral position.

| **Selected on** | Measurement Dataset | Synthetic Dataset |
|---|---|---|
| **Selection Number** | High–level Current [mA] | High–level Current [mA] |
| 1 | 484 | 409 |
| 2 | 451 | 484 |
| 3 | 382 | 444 |
| 4 | 433 | 494 |
| 5 | 352 | 384 |

Table 29: 5 most informative spectral measurement positions selected on the measurement and synthetic dataset.

| **Sample** | **Acetone** [ppmV] | **Ethanol** [ppmV] | **Water** [%V] | **$CO_2$** [%V] |
|---|---|---|---|---|
| 1 | 1.2 | 5 | 0.99 | 1.9 |
| 2 | 0.4 | 5 | 1.7 | 3.2 |
| 3 | 0.8 | 1 | 1.5 | 4 |
| 4 | 0.8 | 5 | 1.6 | 1.9 |
| 5 | 0.4 | 1 | 1.5 | 3.1 |
| 6 | 1.2 | 3 | 0.99 | 3.9 |

Table 30: Concentrations of the additional reduced test set.

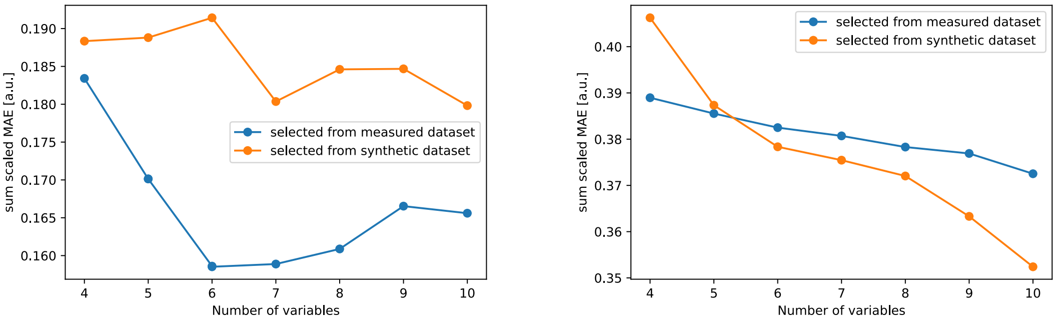


Figure 59: Scaled sum-MAE [a.u.] of a PLS2 trained on the (a) measured and (b) synthetic dataset by number of selected variables and selection mode. In contrast to the (a), the models trained on the synthetic dataset (b) follow the expected downward trend.

The final number of variables is a compromise between decreased performance and a shorter measurement time. For each analyte, at least one measurement point needed to be included. Any additional measurement point could be used to decrease the influence of noise and differentiate better between the three components. To determine the best number of variables, the performance of a PLS 2 for different numbers of variables was performed on the measurement and the synthetic dataset and is visualized in Figure 59. Here, a strong difference between the two training datasets can be seen. While the models trained on the synthetic dataset follow the downward trend expected from theory, the models trained on the measured dataset do not show such a clear trend. This is due to the small size of the measurement dataset, with only 26 samples. When only a few variables are selected per sample, outliers strongly influence the final model. We selected a subset of five variables by keeping in mind the time constraints of the final task. A measurement of five variables can be performed in only five minutes. Through additional engineering, the measurement time could be further reduced. For the evaluation of the reduced model, an additional test dataset measured only at the required spectral positions was acquired. The concentrations are provided in Table 30.

Similar to the previous evaluation in Subsections 7.1.3 and 7.1.4, the same Bayesian hyperparameter tuning as implemented by Weights and Biases [146] for 100 runs was performed for each model, selection basis and training dataset. The parameter range was also kept the same. The measurement models were tuned using a four-fold cross-validation, while the synthetic models were evaluated on a synthetic validation set. The final evaluation took place on the additional measured test dataset. The sum of the scaled MAE similar to the previous evaluation is presented in Figure 60. For this, again, the MAE is scaled by the typical concentration limit of each component, 6 % for water, 5 ppmV for acetone, and 10 ppmV for ethanol, and subsequently summed. From the summed scaled MAE already, a large drop in performance compared to the models evaluated on the full spectrum can be seen. Here, a large percentage of the models show scores above one, which presents a bad performance. The best-performing models are the ensemble models trained on the measurement dataset for both selection modes and trained on the synthetic dataset in combination with the measurement point selection on the synthetic dataset with scores from 0.53 to 0.59. Also among the best performing models are the PLS 1 and 2 trained on the measurement data set with measurement point selection on the measurement dataset with a scaled sum-MAE of 0.56 and 0.5. Finally, the Support Vector Regression (SVR) trained on the measurement dataset with the selection on the synthetic dataset reaches the best overall score of 0.48. For comparison, the best scores of the full spectral models on the measurement test set were 0.17 for the GB model trained on the synthetic dataset and 0.11 for the PLS2 and ANN2 model trained on the measurement dataset as can be seen from Figure 54.

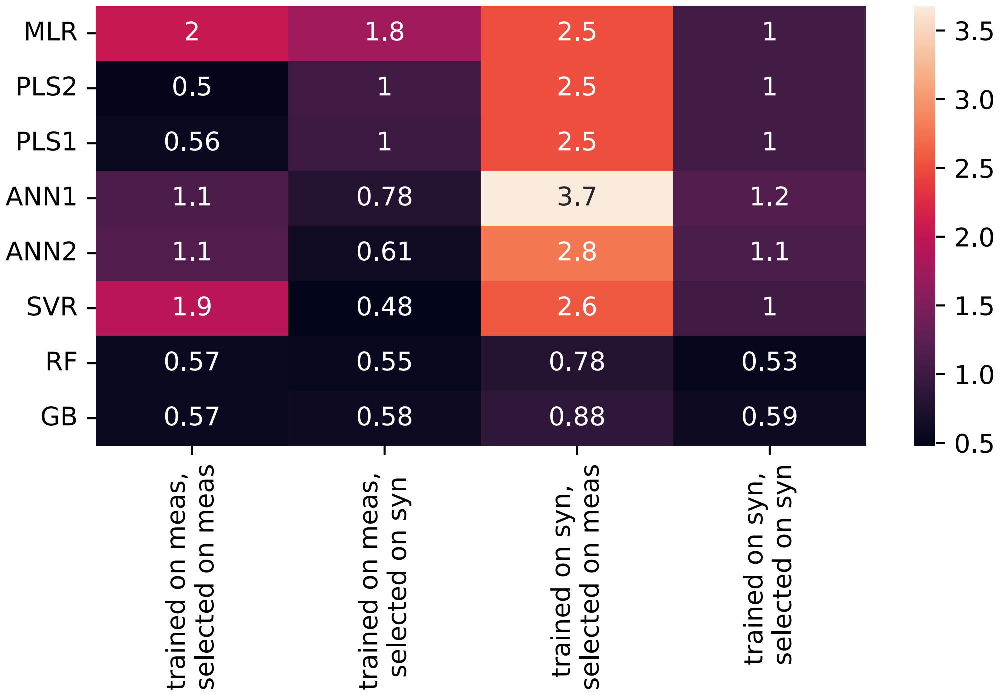


Figure 60: Sum of the scaled-MAE for all models, training datasets, and selection modes evaluated on the additional reduced measurement dataset with five variables.

To evaluate those results more in-depth, the MAE and its standard deviation for water, acetone, and ethanol are summarized in the Tables 31 to 33. The essential overall result visible from Table 33 is that ethanol prediction suffers most from the reduction of variables. The best-performing models can only reach a MAE of around 3 ppmV. This stands in comparison to 0.1 ppmV for the best performing RF model using the full spectral range trained on the augmented measurement dataset presented in the previous section and Table 28. Ethanol has a very shallow slope and was already challenging to quantify from the full spectrum. Hence, its prediction quality suffers most from variable selection. Both other components also suffer from the reduction of measurement points, which is expected. If the entire spectral range is used, multiple points can be considered to corroborate the predicted concentration, while in the reduced set, only five points are available. For the water prediction, the best-performing models on the reduced dataset (ANN2, RF, and GB, all trained on synthetic data, with measurement points selected on measurement data) reach a MAE of around 0.4 %V which is an entire order of magnitude more than the best-performing models reached on the full measurement dataset. On the acetone prediction, the increase in MAE is not as large as in water or ethanol. The best-performing models on the additional reduced measurement test set with only five measurement points in Table 32 show a MAE of around 0.4 ppmV (PLS 1 and 2 trained on measurement data with points selected on measurement data, and SVR trained on measurement data, selected on synthetic data). On the full dataset, the best model reached a MAE of 0.14 ppmV.

| **Model** | **Trained on meas, Selected on meas** | **Trained on meas, Selected on syn** | **Trained on syn, Selected on meas** | **Trained on syn, Selected on syn** |
|---|---|---|---|---|
| MLR | 0.698 ± 0.428 | 0.666 ± 0.455 | 1.26 ± 0.744 | 1.15 ± 0.766 |
| PLS2 | 0.735 ± 0.378 | 0.693 ± 0.372 | 1.26 ± 0.744 | 1.15 ± 0.766 |
| PLS1 | 0.975 ± 0.502 | 0.620 ± 0.311 | 1.26 ± 0.744 | 1.15 ± 0.766 |
| ANN1 | 0.872 ± 0.426 | 0.708 ± 0.337 | 0.708 ± 0.473 | 0.764 ± 0.473 |
| ANN2 | 0.983 ± 0.492 | 0.684 ± 0.398 | 0.383 ± 0.132 | 0.870 ± 0.633 |
| SVR | 1.05 ± 0.536 | 0.597 ± 0.275 | 1.186 ± 0.704 | 0.730 ± 0.528 |
| RF | 0.765 ± 0.368 | 0.695 ± 0.344 | 0.440 ± 0.232 | 0.640 ± 0.371 |
| GB | 0.654 ± 0.359 | 0.896 ± 0.449 | 0.469 ± 0.233 | 0.605 ± 0.371 |

Table 31: MAE on water and its standard deviation [%V] on the additional measurement dataset with only five variables for all training datasets and selection modes.

| Model | Trained on meas, Selected on meas | Trained on meas, Selected on syn | Trained on syn, Selected on meas | Trained on syn, Selected on syn |
|---|---|---|---|---|
| MLR | 1.54 ± 0.280 | 1.33 ± 1.05 | 2.44 ± 0.520 | 1.11 ± 0.799 |
| PLS2 | 0.377 ± 0.275 | 1.270 ± 1.01 | 2.44 ± 0.520 | 1.11 ± 0.799 |
| PLS1 | 0.421 ± 0.320 | 1.035 ± 0.853 | 2.44 ± 0.520 | 1.11 ± 0.799 |
| ANN1 | 0.703 ± 0.418 | 0.889 ± 0.557 | 4.22 ± 1.13 | 1.29 ± 0.959 |
| ANN2 | 0.800 ± 0.427 | 0.579 ± 0.420 | 3.05 ± 0.585 | 1.22 ± 0.817 |
| SVR | 1.87 ± 0.952 | 0.434 ± 0.214 | 2.56 ± 0.550 | 1.10 ± 0.778 |
| RF | 0.652 ± 0.292 | 0.676 ± 0.319 | 0.650 ± 0.261 | 0.679 ± 0.335 |
| GB | 0.765 ± 0.306 | 0.648 ± 0.294 | 0.807 ± 0.251 | 0.726 ± 0.354 |

Table 32: MAE on acetone and its standard deviation [ppmV] on the additional measurement dataset with only five variables for all training datasets and selection modes.

| Model | Trained on meas, Selected on meas | Trained on meas, Selected on syn | Trained on syn, Selected on meas | Trained on syn, Selected on syn |
|---|---|---|---|---|
| MLR | 16.21 ± 4.59 | 13.90 ± 10.89 | 18.02 ± 5.11 | 6.29 ± 3.83 |
| PLS2 | 3.06 ± 1.21 | 6.65 ± 4.77 | 18.02 ± 5.11 | 6.29 ± 3.83 |
| PLS1 | 3.13 ± 1.67 | 6.87 ± 5.15 | 18.02 ± 5.11 | 6.29 ± 3.83 |
| ANN1 | 8.32 ± 4.20 | 4.88 ± 2.50 | 27.18 ± 7.94 | 7.71 ± 5.33 |
| ANN2 | 8.20 ± 4.22 | 3.80 ± 2.20 | 21.01 ± 5.02 | 7.17 ± 4.69 |
| SVR | 13.90 ± 7.44 | 2.93 ± 1.67 | 18.53 ± 5.25 | 6.96 ± 4.31 |
| RF | 3.08 ± 1.79 | 2.98 ± 1.78 | 5.81 ± 1.59 | 2.84 ± 2.30 |
| GB | 3.10 ± 1.63 | 3.05 ± 1.78 | 6.37 ± 1.20 | 3.39 ± 2.30 |

Table 33: MAE on ethanol and its standard deviation [ppmV] on the additional measurement dataset with only five variables for all training datasets and selection modes.

This evaluation on the additional measurement test set with only five measurement points can be summarized to yield the following results:

- **Less measurement points result in less accurate predictions.**
  As expected, the reduction in measurement points has greatly reduced the predictive quality of the models. As pointed out, the decline is worst for ethanol prediction, which is the most difficult to differentiate from the other components. Here, and for the two other components, additional measurement points can be used to reduce noise from the measurement and increase predictive quality.

- **Training on purely synthetic data does not reach comparable performances.**
  As visualized in Figure 60, all models trained on the synthetic dataset, displayed in the two rightmost columns, only show a scaled sum-MAE above one, except for the ensemble methods RF and GB. This bad performance of all not-ensemble methods hints towards a systematic difference in the two datasets, which has already been pointed out in Subsections 7.1.3 and 7.1.4. It also has to be noted that hyperparameter tuning for the two PLS have resulted in a high number of components (15-25), which is too much for only five data points. Ensemble methods like RF and GB alleviate this overfitting and can transfer better to the small measurement dataset.

- **The dataset used for measurement point selection doesn't greatly influence performance:**
  The data basis for the CovSel variable selection was either the measurement or the synthetic dataset. In the resulting models, no selection method notably outperforms the other. A difference is visible for the models trained on the synthetic dataset, which perform much better with the data points selected using the synthetic dataset. But as has already been pointed out using Figure 58, the selected points are in similar domains for most measurement positions. In addition, the broad structures of acetone and ethanol can be detected similarly well from multiple positions. Hence, the selection does not play a prominent role in the final performance.

- **Simple and ensemble methods perform best on this dataset.**
  As visualized in Figure 60, the best performing models are either ensemble methods or a PLS or SVR trained on the measurement dataset. Those models also stand out, as their hyperparameters are well selected to counter overfitting with only three to four components for PLS and a linear kernel for the SVR. As such, again, the linear nature of the dataset already pointed out in the previous chapter is notable, as simple models with few parameters already perform well. Larger, more complex models cannot bring an additional benefit in this case. Only ensemble methods also perform similarly well as they also counter overfitting well.

This concludes the evaluation of synthetic data for a am-PAS setup for human breath analysis. The synthetic data generation has been explained, and the generated synthetic dataset has been used to train different machine learning models. The performance of those models has been compared to the same setup using a small measured training dataset. The results showed that the domain difference between the synthetic and measurement dataset is still too large to apply models trained on the synthetic dataset directly on measured data. Finally, as for medical applications, the measurement time needs to be considered; a selection of only five measurement points has been performed on the synthetic and measurement datasets. While the data basis for the measurement point selection did not greatly influence the result, again, the simple models trained on the measurement dataset and ensemble methods have shown to be the most effective choice. Nevertheless, a strong drop in performance due to the reduction of measurement points was shown. Overall, quantification on the dataset can be solved by mostly simple linear methods, which alleviates the need for large amounts of training data. For such models, the impact of additional synthetic training data is not given.

The following section presents a different dataset created using Wavelength Modulation (wm)-PAS and applied for natural gas or environmental monitoring or and a similar analysis. Here, we have additional non-linear relaxation effects that need to be considered and require more complex machine learning methods.

## 7.2 Photoacoustic Setup for Natural Gas Monitoring

The concept of synthetic photoacoustic spectra creation was also tested on a second dataset to show its broad applicability. This section describes the synthetic data generation and evaluation of machine learning models trained on synthetic and measured data of a wavelength modulated Quartz-Enhanced Photoacoustic Spectroscopy (QEPAS) setup for methane and water detection. It is based on the data published in [208]. After describing the data acquisition setup and its application, the synthetic data generation will be described in Subsection 7.2.1. Here, a particular focus is on the simulation of the wm-PAS signal in contrast to the previous amplitude modulated data and on the integration of the Algorithm to Compute the Collision Based Non-radiative Efficiency and Phase Lag of Energy Relaxation on a Molecular Level (CoNRad) [16] used to estimate the photoacoustic efficiency of the transition. In Subsection 7.2.2 machine learning models trained on measured and synthetically generated data are compared and evaluated in Subsection 7.2.2.

The quantification of methane ($CH_4$) gas is crucial for many applications. Methane is not only one of the main gases that cause the greenhouse effect, which drives climate change [209] but also used in industry and the main constituent of natural gas. As such, its detection over a broad range of concentrations is required. The wm-QEPAS sensor developed by Zifarelli et al. [208] is targeted at methane between 25 and 10 000 ppmV in gas mixtures containing up to 20 000 ppmV $H_2O$. It is based on a Distributed Feedback Laser (DFB)-ICL emitting Mid-infrared (mIR) light between 2987 and 2990 cm$^{-1}$, thus covering two strong $H_2O$ peaks at 2987.5 and 2988.6 cm$^{-1}$ and one strong absorption structure of $CH_4$ from 2988.7 to 2989.2 cm$^{-1}$ as well as a weak $CH_4$ peak at 2987.9 cm$^{-1}$. The absorption cross-section of the two analytes is visualized in Figure 61. The ACS was created with data from the HITRAN database [14] using the Voigt simulation [26] as implemented in hapi [27].

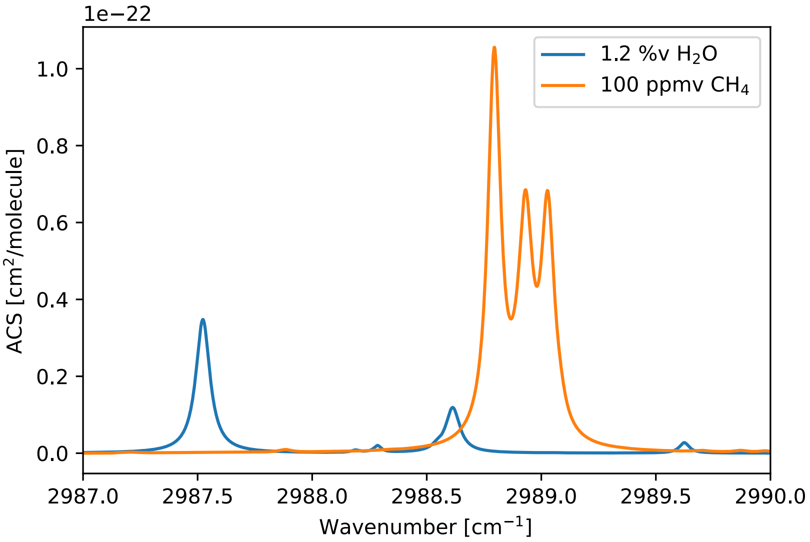


Figure 61: Simulated absorption cross-section of methane and water between 2987-2990 cm$^{-1}$. Two strong $H_2O$ peaks at 2987.5 and 2988.6 cm$^{-1}$ and one strong absorption structure of $CH_4$ from 2988.7 to 2989.2 cm$^{-1}$ as well as a weak $CH_4$ peak at 2987.9 cm$^{-1}$ can be seen.

The measurement setup is schematically depicted in Figure 62. The light source is a DFB-ICL (Thorlabs ID3345HHLH–A) with a central emission wavelength at 2989 cm$^{-1}$ controlled via a Temperature Controller (TEC) to 15 °C. The Laser Driver (LD) is a custom-designed Printed Circuit Board (PCB) controlled by the mainboard. The measurement cell includes a pair of resonator tubes for acoustic resonance amplification and a custom T-shaped Quartz Tuning Fork (QTF) with a resonance frequency of 12 458 Hz and a quality factor of 15 600. The whole cell is controlled to an operating pressure of 400 Torr. Behind the measurement cell, a reference cell containing a certified mixture and a photodiode (Thorlabs PDA07P2) for optical power detection can be found. The photodiode and QTF signal are interpreted and lock-in amplified by a custom Field Programmable Gate Array (FPGA) design on a Red-Pitaya FPGA (STEMlab 125-14). The measurements were performed in 2f-wavelength modulation with a sinusoidal-shaped modulation current. A more in-depth description can be found in [208], including all exact components of the setup.

| Component | Concentrations |
|---|---|
| Methane [ppmV] | 25, 50, 75, 100, 150, 200 |
| Water [%V] | 0.26, 0.5, 0.74, 0.89, 0.98, 1.01, 1.36, 1.54, 1.75, 1.85 |

Table 34: Concentration grid for the QEPAS measurement dataset.

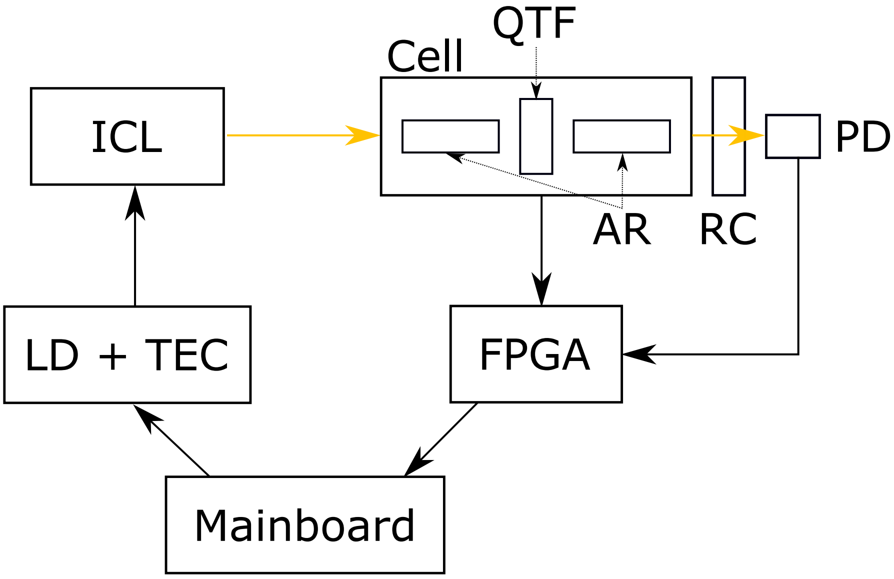


Figure 62: Schematic depiction of the QEPAS measurement setup. The laserdriver (LD) and temperature vontroller (TEC) are controlled by the mainboard. The LD provides the current to the interband cascade laser (ICL) which emits light into the cell. The cell holds two acoustic resonators (AR) and the quartz tuning fork (QTF). The cell is followed by a reference cell (RC) and a photo diode (PD). The signals from the PD and QTF are amplified by a field programmable gate array (FPGA) and sent back to the mainboard.

A total of 60 samples were acquired with the sensor in spectral scan mode. Each methane concentration provided in Table 34 was combined with each water concentration, yielding a total of 60 measurements. Each spectrum consists of 537 data points with a spectral resolution of about 0.0036 cm$^{-1}$ covering the range from 2987.0 to 2989.5 cm$^{-1}$.

**Special challenges of this dataset include:**

- Strong change in the photoacoustic efficiency of $CH_4$ in relation to the water concentration.
- Saturation of the sensor at the $CH_4$ structure between 2988.7 to 2989.2 cm$^{-1}$ for high methane concentrations
- Double resonant features from the QEPAS setup

While the first notion is currently included in the modeling setup, saturation, and double resonance are neglected in the synthetic data generation. The saturation of the methane peak does not appear in the current setup, and additional measurements by Zifarelli et al. with this effect were not employed in this work. This alleviates the need for saturation integration in the setup. Nevertheless, the modeling chain can easily include this saturation. It can be integrated at the first step of the modeling chain - the generation of the absorption spectra. Including the double resonance features from the measured resonance values would be an excellent extension of the current framework and an exciting avenue for further research.

The data and additional information on the sensing setup were kindly shared by Zifarelli et al. for this work [208].

### 7.2.1 Synthetic Data Creation

To generate a physically grounded wm-PAS spectrum in this wavelength range, a similar procedure as for the previously described am-PAS spectrum was followed. From the steps described in Section 7, steps one to seven and step nine were employed. The simulation of the background signal was not considered for a Second Harmonic (2f) measurement, as no background is expected. In a 2f measurement, the detection frequency is set to the second harmonic of the modulation frequency. As now a wm signal is to be simulated, the signal is no longer directly connected to the height of the absorption coefficient of the mixture but to its slope over the modulation depth. Hence, in step six, the simulation of this signal is required. This corresponds broadly to the derivative of the signal, but to allow the modulation depth to be integrated into the simulation setup and best model the actual signal generation, a different approach employing a Fast-Fourier Transformation (FFT) is chosen. For this, the am PAS signal for a fine grid of wavelengths around the center current is simulated. Those simulations are concatenated following a simplified modulation of the input current. Hence, the signal in the time domain is generated. To convert this signal to the spectral domain, a FFT is employed, similarly to the signal acquisition in the LIA.

The signal at each center current $I_c$ around which the signal was modulated was computed following Equation 89. $q(I_c)$ corresponds to a signal amplification factor, the Fourier transform $\mathcal{F}_t$ is taken over the expected signal $s(I_c, m, t)$ at each time $t$ for a given modulation depth $m$ at center current $I_c$. This expected signal is computed similarly to the amplitude modulated signal following Equation 90 where $P_{\tilde{\nu}}(I(I_c, m, t))$ corresponds to the ICL output spectrum at the input current $I$ which is computed assuming a linear current increase and decrease over the full modulation depth $m$ at center current $I_c$. This signal is convoluted with the simulated mixture absorption spectra $\sum_c A_{\tilde{\nu}c}\epsilon$, which now integrates the photoacoustic efficiency $\epsilon$ to scale each component's simulated spectra $A_{\tilde{\nu}c}$ before summation.

$$p_a''(I_c) = q(I_c)\mathcal{F}_t\left(s\left(I_c, m, t\right)\right) \tag{89}$$

$$s(I_c, m, t) = P_{\tilde{\nu}}(I(I_c, m, t)) * \sum_c A_{\tilde{\nu}c}\epsilon \tag{90}$$

The absorption spectra of water and methane $A_{\tilde{\nu}c}$ were simulated using HITRAN line data [14] and a Voigt profile [26] in hapi [27]. The temperature was set to 28 °C, and the pressure was 400 Torr. Those spectra were generated as a concentration grid from 0.2 - 2.1 %V for water and 10 to 250 ppmV for methane. For the second step of the simulation chain, the photoacoustic efficiency $\epsilon$ needs to be estimated for both molecules. Prior research has shown a higher water concentration to promote the photoacoustic efficiency of methane [16], [53]. The photoacoustic efficiency $\epsilon$ of methane at this excitation wavenumber for each water concentration was computed following the CoNRad algorithm [16], [208]. The result is visualized in Figure 63 and shows a nonlinear influence of the water concentration. The simulated methane spectrum was multiplied by the efficiency factor. It was considered one for water in general. Following the third step, the methane and water spectra were added to generate the mixture spectrum $\sum_c A_{\tilde{\nu}c}\epsilon$.

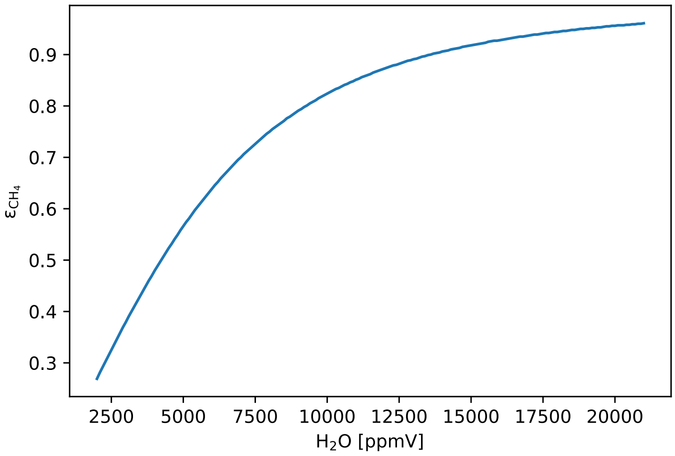
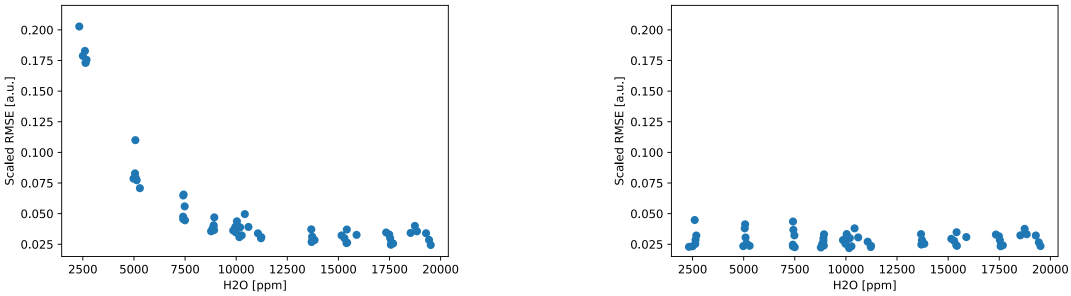


Figure 63: Photoacoustic efficiency of the methane transition by increasing water concentration.



(a) Without integration of the photoacoustic efficiency.

(b) With the integration of the photoacoustic efficiency.

Figure 64: Scaled RMSE of the synthetic spectra compared to the measurement by $H_2O$ concentration. a) without and b) with the integration of photoacoustic efficiency. A high error for low water concentration without the integration of the photoacoustic efficiency can be seen in subfigure a).

The importance of integrating this efficiency can also be seen in Figure 64. Here, the scaled RMSE is depicted in relation to the sample's $H_2O$ concentration. The residual used to compute the RMSE between the synthetic and measured spectra has been scaled by the maximum value of each measured spectrum to be able to compare the different concentrations and signal heights. As the photoacoustic efficiency approaches 100 % for higher water concentrations, the effect on the reconstruction error diminishes for those higher concentrations in Figure 64a.

The simulation of the laser output spectrum as the fourth step was simplified compared to the same step in Subsection 7.1.1. The output of the ICL used here was considered to have a Lorentz-shaped profile, as indicated by the datasheet. Therefore, only three parameters were used to simulate this output following the Lorentz distribution $L$ - also known as the Cauchy distribution and presented in Equation 91. The center of the output was fitted with a linear and constant component $b_0, b_1$, while the scaling of the function was estimated as a constant $b_2$. As a different research group generated the data, no spectral scans of the laser output were available to fit those parameters directly and confirm the laser output provided by the datasheet. Instead, they were fitted directly on the photoacoustic spectra.

$$L(\tilde{\nu}, b_{0-2}, I_c) = \frac{1}{\pi(1 + ((\tilde{\nu} - (b_0 I_c + b_1))/b_2)^2} \frac{1}{b_2} \tag{91}$$

Finally, in step five, the convolution of the laser output spectrum and the absorption spectrum is performed as indicated in Equation 90. During each measurement point in wm-PAS, the laser output is modulated around the center wavelength to generate the photoacoustic signal. This had to be considered during the simulation. First, the output of the laser during one complete modulation cycle is simulated. The laser modulation is simplified from the actual sine-shaped signal to a triangular shape. This change is justified, as the laser transfer function shows substantial deviations to the actual signal for such fast modulations. The laser current modulation was simulated with a computation step of 0.01 mA for an entire modulation cycle. As this modulation is constantly applied, the signal simulated for one modulation cycle was extended fifty times for computational stability. Next, a FFT was used on this time-based signal to convert it to the spectral domain as indicated in Equation 89. A high discrepancy between the applied modulation depth of 2.1 mA and the one estimated by simulation of 0.52 mA can be explained by two factors. Firstly, the simplification of the sine-modulation function to a triangular signal, and secondly, the actual laser transfer function is not expected to follow the one indicated in the datasheet for such a fast modulation. The current system does not perfectly model the rapid modulation, but the impact is expected to be small due to the following computational steps. This estimation of the second order PAS signal is repeated for each center driving current $I_c$ of the ICL to generate the entire spectrum.

The eighth step, the estimation of the background signal does not apply to the wm-PAS second-order signal. But the amplification is estimated as a combination of a linear and constant component as indicated in Equation 92. This leads to a final signal estimation.

$$q(I_c) = d_0 I_c + d_1 \tag{92}$$

The simulation parameters were fitted using 40 spectra from the measurement data, which corresponds to 66 % of the available data. A combination of least squares and Nelder Mead fitting was employed to reach the best performance. The final setup got an overall RMSE of 5.733 [a.u.] and a scaled RMSE of 0.0196 [a.u.], where the residual is scaled by the maximum of the measured spectrum. A correlation with the water concentration is no longer visible, as presented in Figure 64b. Figure 65 shows a typical example of the synthetic data. The main error in simulations stems from the peak differences for methane, visible at 118 and 122 mA.

Finally, the variance of the parameters over the evaluation set is estimated using the covariance matrix. As described in Section 7.1.1, this step can only be considered trustworthy if no strong correlations between the parameters exist. This was verified using MCMC estimation [205], where no strong correlations were visible. Table 35 presents the fitted values and estimated relative variance.

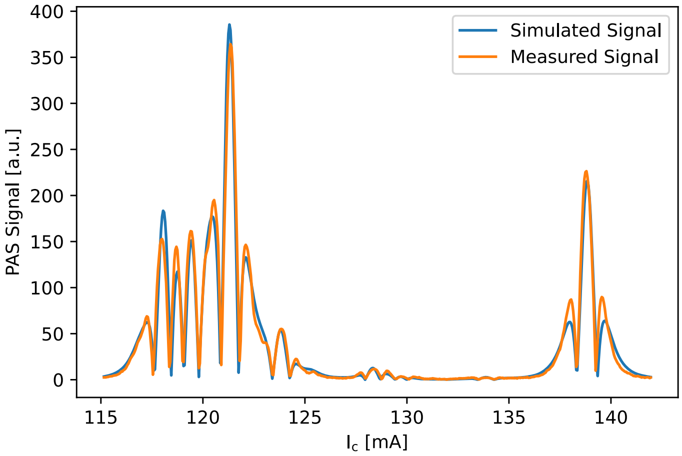


Figure 65: Example comparison of a simulated and measured wm-PAS signal at 1.75 %V water and 100 ppmV $CH_4$. The main error stems from differences in $CH_4$ peak height at 118 and 122 mA.

| **Parameter** | **Value** | **Description** | **Estimated relative variance** |
|---|---|---|---|
| $b_0$ | 0.013081 | HWHM of the ICL | 2.2590e-05 |
| $b_1$ | -0.072573 | instrument transfer linear with $I_c$ | 1.8486e-07 |
| $b_2$ | 2997.60 | instrument transfer constant | 2.2616e-05 |
| $c_0$ | 0.52323 | modulation depth | 5.6290e-10 |
| $d_0$ | 0.40153 | PAS signal transfer scaling linear with $I_c$ | 3.3144e-03 |
| $d_1$ | 436.0008 | PAS signal transfer scaling constant | 0.54988 |

Table 35: Fitted parameters for the wm-PAS ICL synthetic data generation, their values, estimated variance, and physical correspondence.

The presented method for synthetic data generation of the wm-PAS ICL data was employed to generate a synthetic training dataset. Here, the parameters were sampled from a normal distribution of one standard deviation, and a grid of concentrations was generated. The methane concentration was varied from 10 to 250 ppmV in steps of 5 ppmV, while the water concentration was increased from 0.2 to 2.1 %V in steps of 0.1 %V. This leads to a total of 9120 combinations and generated spectra, which were used to train and evaluate machine learning models, as presented in the following section.

The quality of this wm-PAS synthetic data is below the one presented in Subsection 7.1.1 as two additional simulation steps are necessary to integrate the photoacoustic efficiency and the wavelength modulation. In addition, the finer structure of methane poses a greater challenge. Those two points still offer room for improvement and future research.

### 7.2.2 Machine Learning with Synthetic Data

The same machine learning models and learning regimes as in Subsection 7.1.2 were employed to provide a comparable evaluation as in the previous chapter. Again, no deep learning methods were used, even though a greater impact of those approaches is expected, as the spectral structures are finer compared to the am-PAS application, and non-spectral, non-linear effects are expected. The algorithms employed were:

- Multi Linear Regression (MLR)
- Partial Least Squares (PLS) 1
- Partial Least Squares (PLS) 2
- Support Vector Regression (SVR)
- Artificial Neural Network (ANN) 1 (one layer)
- Artificial Neural Network (ANN) 2 (two layers)
- Random Forest (RF)
- Gradient Boosting (GB)

The synthetic dataset, including the simulated measurement deviations, was created as described in Subsection 7.2.1. The measurement set was used with the same cross-validation setup as presented in [208], which uses 50 % of the data in five balanced cross-validations. This mode of evaluation does not employ an independent test set nor double cross-validation but was chosen to remain comparable with the work of Zifarelli et al. despite those concerns. All five cross-validation folds were combined in one validation to evaluate the models trained on the synthetic dataset. The remaining 50 % of the measurement data was used to tune the hyperparameters of the machine-learning models. All models were subject to hyperparameter tuning for 200 evaluations with Bayesian optimization as implemented by Weights and Biases [146]. The tuning range of all hyperparameters is kept the same as presented in Subsection 7.1.2 provided in Tables 18 to 22 with the addition of a noise parameter to scale the gaussian noise added to the synthetic spectra which ranged from 0-1 in a uniform distribution. All spectra were preprocessed by subtracting the mean spectrum of the measurement training data and dividing by the mean standard deviation of the training spectra over the entire spectral range. All models were implemented and fitted using the Scikit-learn library [147].

For the evaluation, the MAPE score is used, which is computed as $MAPE = \frac{1}{n}\sum_i^n (|y - \tilde{y}|)/\tilde{y}$ with $y$ as the prediction and $\tilde{y}$ as the target concentration, over all samples $n$. Lower values indicate better performance. This score and its advantages and disadvantages have also been described in Subsection 5.5.5 and were also used by Zifarelli et al. [208].

The evaluation on the synthetic test dataset is displayed in Figure 66 and 67 for water and methane. The results of the models trained on the synthetic dataset are presented in the left columns, while the right column shows the results of the models trained on the measurement dataset. The models are listed on the y axis. The performance is indicated in the grid and color coded. As expected, the models trained on the measurement dataset performed worse than those trained on the synthetic dataset for both components and all models. This is because the synthetic models were trained on a dataset from the same distribution, while the measurement models were trained on the smaller measured dataset. The score difference indicates a strong domain gap still persisting for the synthetic dataset. This domain gap persists for the water, as well as the methane prediction. When focusing on water presented in Figure 66, the prediction is easier for the model as no non-spectral effect influences the absorption, resulting in a lower overall MAPE. All models reach MAPE scores below 0.1, corresponding to an error of below 10 %.

Inspecting the results on methane presented in Figure 67 closer, we can see a different picture for models trained on the synthetic and the measurement dataset. For the synthetic models in the left column, we see the simpler algorithms like MLR and PLS performing worse on the methane prediction with a MAPE of around 0.22 than more complex models hinting towards the non-linearity of the underlying process. The same is visible for the simple models trained on the measurement dataset. The ensemble methods RF and GB trained on the synthetic dataset reach the best performance on methane prediction with a MAPE of only 0.02, which corresponds to a 2 % error on the prediction.

Combining the results from both components' evaluations, for water prediction, the ability of the model trained on the measurement dataset to transfer to the synthetic dataset is greater, leading to lower MAPE scores of only 0.06 - 0.09. The very low MAPE score of the MLR model of 0.06 highlights the linear predictability of water in comparison to methane, where the same model has a MAPE score of 1.29, which corresponds to an error of 127 % on the prediction.

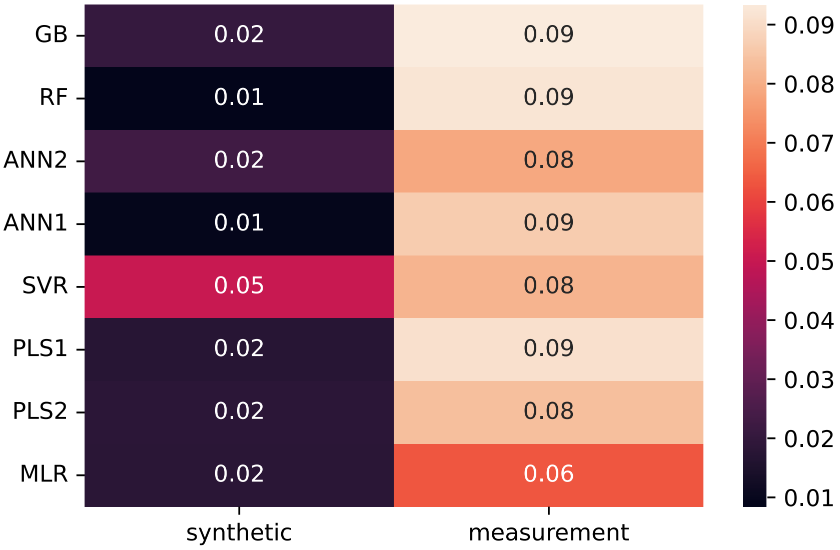


Figure 66: Water MAPE [a.u.] of the models on the synthetic test set. Lower scores indicate better performance.

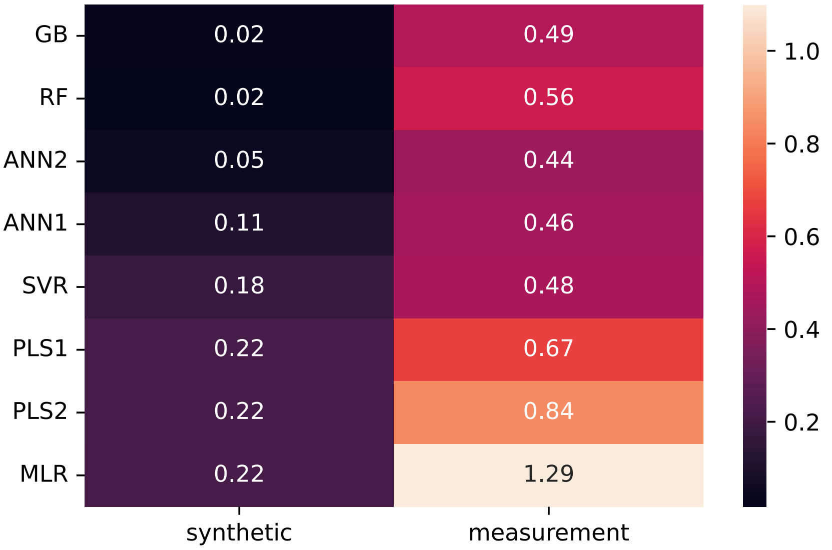


Figure 67: Methane MAPE [a.u.] of the models on the synthetic test set. Lower scores indicate better performance.

When the models are evaluated on the measurement test set as presented in Figures 68 and 69, the inverse effect becomes visible as the models trained on the measurement data now perform better. This is expected, as now the models trained on the measurement dataset were trained on data from the same distribution, and even their hyperparameters tuned towards this dataset. Again, the MAPE of water and methane are visualized as a heatmap in Figures 69 and 68.

On the water prediction, shown in Figure 69, the models trained on the synthetic data perform worse with 0.03 - 0.06 MAPE, corresponding to a 3-6 % error in the prediction. In comparison, the models trained on the measurement dataset show a MAPE of 0.004 to 0.03.

But on the methane prediction, presented in Figure 68, again, a more complex picture emerges. Here, the simple models trained on the measurement dataset (PLS and MLR) reach MAPE scores of more than 0.14 to 0.19 while the more complex models (GB, ANN 2, and SVR) achieve lower 7-8 % prediction errors. For the synthetic models, a similar trend can be seen with the simple models (MLR and PLS) at 26 or 27% prediction error. The more complex models (ANN2, GB, and RF) improve upon this, reaching a MAPE score of 0.07 to 0.09.

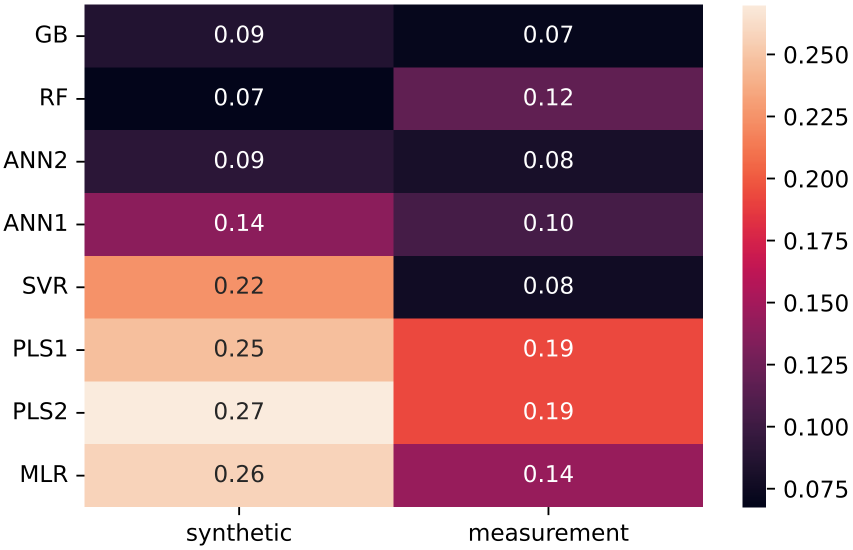


Figure 68: Methane MAPE [a.u.] of the models on the measurement test set. Lower scores indicate better performance.

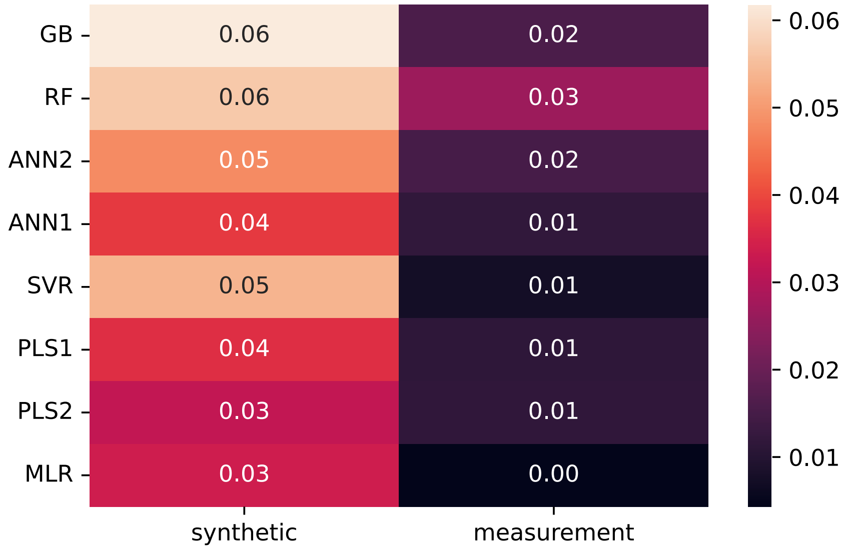


Figure 69: Water MAPE [a.u.] of the models on the measurement test set. Lower scores indicate better performance.

The worse performance of the synthetic models on the water predictions can most likely be alleviated by a fine-tuning step or a different type of domain transfer, as it can be well solved by MLR trained on the measurement dataset. On the other hand, the main analyte methane is more complex to predict due to the non-spectral effects affecting the height of the methane structures. To improve the synthetic models further, either an improvement of the synthetic data generation or an additional fine-tuning of the models trained on the synthetic dataset can be used.

Those results, especially on methane prediction, show that machine learning methods trained on a synthetic dataset can generalize to measured datasets and reach similar performances. In this case, a comparatively large amount of measurements, evenly spread on the concentration grid, was available to train a comparable model on the measurement dataset. The approach to simulate synthetic PAS spectra can be used to alleviate the need to create such large datasets.

It must be emphasized that the results presented here are not directly comparable to those presented in [208]. The models presented in this work have been trained solely on the PAS magnitude spectrum. In contrast, the approach proposed by Zifarelli et al. [208] uses a concatenation of the X and Y signals. They reach a MAPE score - indicated as Average Relative Error of Prediction (AREP) in their work - of 0.046 on methane. The combination of X and Y are the real and imaginary part of the complex PAS signal, while the magnitude corresponds to the distance in a polar coordinate system. The X and Y signals are displayed in Figure 70 and the computed magnitude and phase signal of the same measurement in Figure 71. The phase information is not contained in the simulated data, which is available from the X and Y concatenation used by Zifarelli et al., which leads to the decrease in methane prediction performance. This phase information is vital for their methane prediction as a PLS trained solely on the signal magnitude as presented in this work only reaches a MAPE score of 0.19. Their work was reproduced with the same training setup as used in this work for verification.

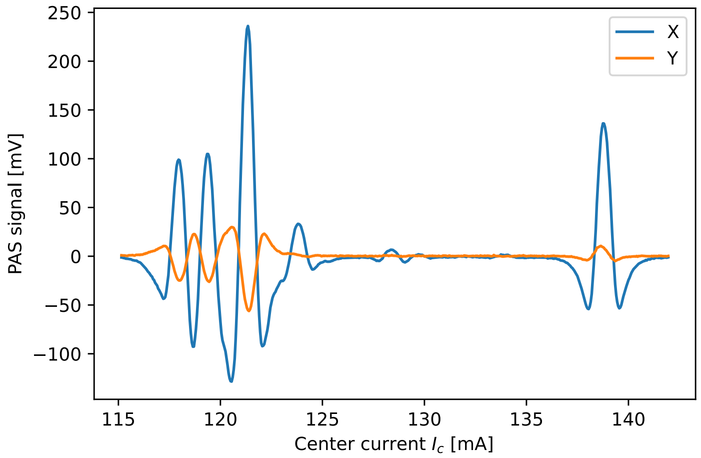


Figure 70: X and Y component of the complex wm-PAS signal at 11 000 ppmV water and 75 ppmV methane.

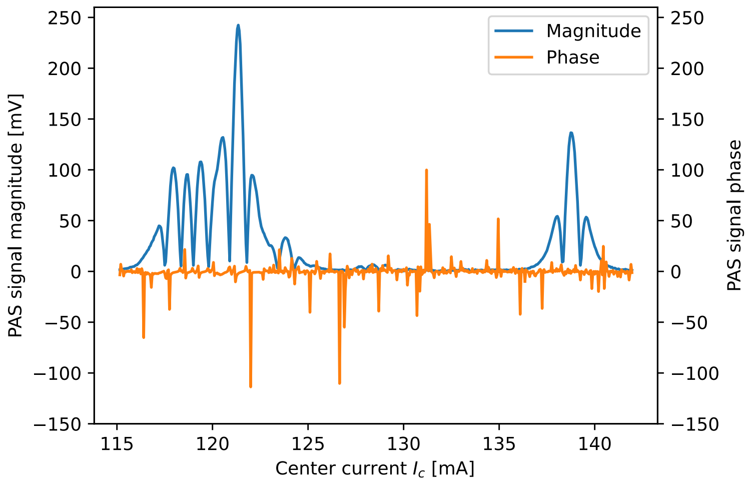


Figure 71: Magnitude and phase component of the complex wm-PAS signal at 11 000 ppmV water and 75 ppmV methane.

This strong impact of the phase information on the prediction quality of a machine learning system highlights the importance of a better understanding of the phase information in the PAS measurement setup, which can also enable a simulation of this signal. The CoNRad algorithm can predict the phase shift from the relaxation process at the peak position but needs to be expanded to the full spectrum for further use [16]. This is, in itself, an exciting topic for future research.

Overall, three main conclusions can be drawn from the machine learning experiments presented in this section:

- **A domain gap between the synthetic and measured dataset remains.**
  The models trained on the synthetic dataset perform worse on the measurement test set, except for complex machine learning methods like ANNs and ensemble methods on methane. The worse performance of the synthetic models on water concentration prediction highlights a remaining domain gap, as water prediction is a linear task. The problems in synthetic signal generation for wm-PAS have already been highlighted in the previous Subsection 7.2.1. Another way to eliminate those domain differences effects is fine-tuning or training the models on a mixture of synthetic and measured data as presented in [181].

- **Synthetic data can reach comparable results on complex tasks.**
  In contrast to the water prediction and the analysis of the am PAS QCL setup presented in Section 7.1, the models trained on synthetic data for methane concentration prediction in this chapter reached a comparable score to the models trained on the measurement dataset even without an additional domain adaptation step. The methane prediction on the wm-PAS dataset presented here is more complex than the other tasks, as a non-linear relaxation effect affects the spectrum. The good result here points to a suitable application area of synthetic PAS training data: Complex non-linear prediction tasks that simple or linear machine learning methods cannot solve.

- **Simple (linear) machine learning models fail to predict methane concentration from the PAS magnitude spectra.**
  The PLS methods applied in this work and also by Zifarelli et al. [208] remain linear as they do not use additional kernel methods (see Chapter 5). Therefore, they don't show good predictive performance for methane, independent of whether they were trained on the synthetic or measured dataset. The result of 0.046 MAPE scores employing a PLS regression from the photoacoustic X and Y signal, therefore, indicates that a linear relationship between the methane concentration and the combined X and Y signal (hence the phase information) exists. This relationship has been exploited in their work to reach the good predictive performance. This finding also highlights the importance of a further study of the photoacoustic phase information by sensor developers as another way to circumvent the non-linear relaxation efficiency of methane in this gas mixture.

The results achieved with different machine learning models trained on synthetic and measured PAS data have been presented and evaluated. Even though a domain difference between the two signals still remains, the task presented here points towards future application areas for synthetic PAS data in more demanding environments where non-linear effects hinder the application of simpler models that can already be fitted with a few measurement examples.

The generation of a synthetic photoacoustic signal from a few measurements and synthetic absorption spectra has been presented for am as well as wm PAS. In the first example, a PAS setup for breath analysis has been presented. The synthetic data generation of an am PAS signal is described, and the synthetic data generated is used to train eight different machine learning models. The same models were trained using a measured or augmented measurement dataset for comparison. The direct performance of the models trained purely on synthetic data did not reach comparable performance to the models trained on the measurement or augmented measurement dataset. When variable selection is employed they can reach en par performances. The risk of overfitting on measurement artifacts significantly increases with variable selection, which might be alleviated by the use of synthetic data.

The second example presents a wm-PAS setup for methane detection under varying humidity conditions. Again, the synthetic data generation from a few measurement samples is described, including the modeling of the wm signal and integration of the CoNRad algorithm for photoacoustic efficiency. A similar machine learning setup as for the am-PAS setup is presented with models trained either on the synthetic or the measured dataset. The results are also compared to the original results published by Zifarelli et al. [208]. The models trained on purely synthetic data reached a comparable performance of methane prediction on the measurement dataset to the models trained directly on the measured dataset. This points towards a future application for synthetic photoacoustic spectra in more challenging scenarios where non-linearities hinder the application of simple machine learning methods.

In the following final chapter, the findings from this thesis will be summarized, and conclusions will be drawn. Finally, areas for future work are outlined.

*Part IV*

# Conclusion

# 8 Summary

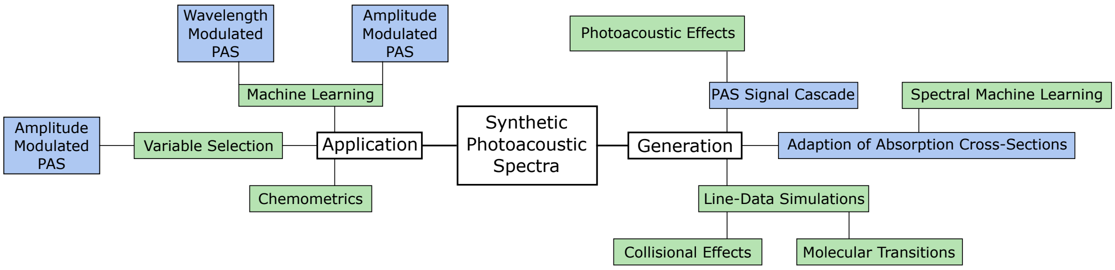


Figure 72: Overview of the topics covered in this thesis. Blue shaded squares correspond to original contributions.

Figure 72 aims to provide a quick overview of the topics covered in this thesis and the original contributions made.

The generation and application of synthetic photoacoustic spectra were explored. To generate those synthetic spectra, a way to reproduce the **photoacoustic signal cascade** is needed, which has been introduced by the author in Chapter 7. This signal modeling cascade transforms absorption spectra into photoacoustic spectra. The method fits a limited number of modeling parameters to a small measured dataset and estimates the variance of those physically grounded parameters. This signal cascade is based on knowledge of the **photoacoustic effects** taking place during signal generation, previously reviewed and described in Chapter 3. The modeling is based on existing absorption spectra, which can either be modeled from **line-data simulations** or converted from measured absorption cross-sections. The simulation from line data is elaborated on in Chapter 4. The underlying physical effects, mainly **molecular transitions**, which form the basis of absorption and absorption line data, are presented in Chapter 2. The **collisional effects** which govern the line shape to be modeled in gas spectroscopy are reviewed in Section 2.3. Since line data is not applicable for all molecules of interest due to the limitations of the current modeling systems, alternative sources of absorption cross-sections are outlined in Sections 4.5 and 4.6. The most common other source of absorption cross-sections is the use of measured data. As those often do not correspond to the environmental configurations of the application, a novel deep learning approach to **adapt absorption cross-sections** to other environmental parameters is presented in Chapter 6.

The approach covered in this work relies on training a deep-learning neural network on a set of simulated absorption cross-sections. From this simulated dataset, the adaption of the spectrum to a higher pressure configuration is learned. The model is evaluated on actual measured cross-section spectra from a molecule unknown to the model. The results perform well but remain less accurate than often used estimated pseudo-line lists. Nevertheless, the deep learning approach can already be applied when less knowledge of the molecule in question is available. The presented approach is extended for continuous pressure adaptation, and its limitations concerning temperature adaption are discussed. For the evaluation of the performance of this approach, a review of **spectral machine learning** is provided in Section 5.7.

Synthetic photoacoustic spectra are employed to counter data scarcity in quantitative spectroscopic applications. To provide an overview of current applications of **machine learning** in (photoacoustic) gas absorption spectroscopy, the field is reviewed in Chapter 5, and the machine learning techniques used in this work are presented. Particular focus is placed on reviewing and integrating the vast research field of **chemometrics** into the machine learning landscape in Section 5.5.

Synthetically generated photoacoustic training data is compared to measured datasets in two applications to showcase its use. A first, extensive study has been performed on an **amplitude modulated photoacoustic** dataset for the quantification of acetone, ethanol, and water in human breath exhale. This is presented in Section 7.1. The data generation is explained in detail in Subsection 7.1.1, and multiple different machine learning approaches are compared on the various training regimes in Subsection 7.1.2 to 7.1.4. This provides an extensive study on the effects of synthetic data, augmentation, and small measured datasets over a range of different machine learning algorithms. The results show a remaining large domain gap that hinders the direct application of models trained purely on synthetic data. The same measurement and synthetic dataset are also used for variable reduction to shorten the measurement time of the final application. For the full and reduced datasets, simple partial least squares methods trained on the measurement dataset and ensemble methods trained on either the synthetic or measured dataset perform best. This is due to the high linearity of the dataset, which does not require more complex methods to predict the concentration.

A second application was evaluated to corroborate the results presented on the amplitude-modulated photoacoustic dataset. The results of the synthetic data generation and their use as a dataset for machine learning are presented in Section 7.2. Here, methane and water are quantified for a natural gas monitoring application. The photoacoustic signal cascade for **wavelength modulated photoacoustic** simulations is presented, and the integration of non-spectral relaxation effects is introduced in Subsection 7.2.1. A study of the impact of the underlying training dataset, synthetic or measured, is presented on a range of machine learning algorithms in Subsection 7.2.2. The results show that non-linear machine learning models like Artificial Neural Networks (ANNs) and ensemble methods trained on purely synthetic data reach comparable performances on methane concentration prediction. Nevertheless, further improvement of the wavelength-modulated modeling is needed, especially the simulation of the phase information, which is an exciting avenue for future research.

# 9 Future Work

This chapter outlines three directions for future research building on the results and methods presented within this thesis. The selected topics are not encompassing all possible ways to build upon the findings and theory presented within this work. Still, the author believes those three directions to be the most promising. They were not performed within this thesis due to time constraints.

**Integration of molecular information for spectral adaption.**
The limitations of the deep learning network for spectral adaption presented in Chapter 6 stem from the missing knowledge of the model about the molecule and molecular transitions in question. Therefore, integrating this knowledge into the model could significantly improve its application area. Molecular fingerprints, commonly used to encode chemical information for neural networks, are a broad field of research. They can be combined with knowledge about ab initio simulations for absorption cross-sections. The first steps towards this extension are presented in Section 6.5.

**Integration of phase simulation into the synthetic photoacoustic data generation.**
The phase information in photoacoustic spectroscopy can be used to circumvent non-linear relaxation effects as outlined in the conclusion of Chapter 7.2.2. Its integration into the simulation toolchain could significantly improve the application of synthetic photoacoustic data and provide further understanding of the molecular interactions in the gas matrix. A start towards the prediction of the phase information of specific transitions has been integrated into the Algorithm to Compute the Collision Based Non-radiative Efficiency and Phase Lag of Energy Relaxation on a Molecular Level (CoNRad) algorithm by Müller et al. [16], which could be extended with the synthetic data generation approach from Chapter 7.

**Application of synthetic data and deep learning models for complex photoacoustic spectra interpretation.**
The results presented in Section 7.2.2 showed that synthetic data and more complex machine learning approaches should only be applied when simple linear methods fail to interpret the photoacoustic signal correctly. Therefore, the applications that can benefit most from synthetic photoacoustic spectra lie in even more complex signals and probably a larger scanned wavelength range. For example, those larger wavelength ranges can be reached by dual-comb photoacoustic spectroscopy, which has recently been introduced by Friedlein et al. [210]. The combination of synthetic data with Convolutional Neural Network (CNN) [99] or Transformer architectures [100] for spectra interpretation are believed to hold great potential, as those approaches require substantially larger datasets.

# 10 Conclusion

This thesis provides an excellent resource for both spectroscopists and machine learning experts interested in simulating (photoacoustic) spectral data and its application in machine learning for gas component quantification. In addition to the theory of absorption and especially photoacoustic spectroscopy, a comprehensive review of machine learning methods commonly used in spectroscopy with a focus on gas spectroscopy is provided. The current possibilities in gas spectral simulation are presented, and an additional deep-learning method to extend measured absorption cross-sections to other environmental configurations is developed, evaluated, and discussed. Finally, the simulation of Photoacoustic Spectroscopy (PAS) spectral measurements is presented, including amplitude and wavelength-modulated measurement setups and the incorporation of non-spectral interference. Those synthetic training datasets are compared to their measured counterparts in two setups over a range of commonly used machine learning algorithms. In an Amplitude Modulation (am) PAS sensor setup for joint acetone, ethanol, and water quantification in human breath exhale conditions, variable selection with Covariance Selection (CovSel) was performed. Due to the remaining domain gap between measured and synthetic data, they do not perform as well as their measured or augmented counterparts. The dataset used for variable selection only plays a minor role in this application. Overall, the linear nature of this dataset alleviates the need for more complex machine learning methods; hence, the benefit of synthetic data is only marginal. In a Wavelength Modulation (wm) Quartz-Enhanced Photoacoustic Spectroscopy (QEPAS) setup for methane detection under substantial humidity changes, synthetic data was generated with an incorporation of the Algorithm to Compute the Collision Based Non-radiative Efficiency and Phase Lag of Energy Relaxation on a Molecular Level (CoNRad) algorithm to account for non-spectral interference. Again, a comparison of the measured and synthetic dataset as a training basis for different machine learning algorithms was performed, resulting in en-par performances for methane prediction of each best model. Those results show the applicability of the proposed PAS signal simulation approach to generate synthetic training data, which can improve algorithms for gas quantification in PAS, especially when non-linear effects need to be included.

The main contributions of this work are:

- **Transfer of pressure changes on absorption cross-sections via deep learning**
  In Chapter 6, the leading hypothesis is that a deep learning model trained on environmental changes on simulated absorption cross-section can transfer to other, unknown molecules for which simulation is not yet possible. This hypothesis was evaluated using a Convolutional Neural Network (CNN) including residual and squeeze and excite blocks, which was trained on cross-sections simulated using High-Resolution Transmission Molecular Absorption Database (HITRAN) and hapi. The evaluation was conducted on measured absorption cross-section of $ClONO_2$. The model transferred well for pressure increases but has substantial limitations in the case of pressure decreases and temperature changes. Those limitations have been explained, and possible extensions, like the inclusion of molecular fingerprints, have been proposed.

- **Creation of synthetic photoacoustic spectra**
  In Chapter 7, the adaption process of measured or simulated absorption cross-section to correspond to a measured PAS signal is described. This multi-step adaption process can be used for wm as well as for am PAS data and only requires a few measurements. It follows the physical signal cascade of the PAS signal generation, allowing to incorporate known deviations of the measurement setup. Thus, it can easily be used to augment the created synthetic data. This approach has been used in Sections 7.1.1 and 7.2.1 for a Quantum Cascade Laser (QCL) based am PAS setup as well as for an Interband Cascade Laser (ICL) based wm QEPAS setup. A stronger deviation from the measured signal was observed for the simulated wm data. Therefore, the simulation of the wm process has the highest potential for further improvements. In addition, the current limitations of synthetic data generation, which, for example, do not incorporate phase information, provide a promising avenue for future research.

- **Use of synthetic data to overcome data scarcity in machine learning for photoacoustic spectroscopy**
  While data scarcity is commonly known as one of the main hindrances to applying more complex machine learning models, synthetic data has only recently been used in gas spectroscopic applications and not yet in PAS spectroscopy. An extensive study of the effect of synthetic compared to measured training data on different machine learning algorithms is presented in Chapter 7. Two experiments have been performed on different PAS setups in Sections 7.1.2 and 7.2.2. While the domain gap between synthetic and measured data presents a strong challenge for the transfer of models trained on purely synthetic data, this domain gap can be overcome in the future by domain adaption methods. Those ways to reduce the domain gap, for example, fine-tuning are presented and discussed. Applying synthetic data for machine learning in photoacoustic spectroscopy shows the greatest potential in complex predictive tasks where non-linearities forbid the use of simple linear methods.

Those contributions can be applied in sensor development for multi-gas sensing in complex scenarios, thus allowing the conceptualization and development of novel (photoacoustic) absorption gas sensing devices. Adapting measured absorption spectra to other environmental configurations allows a more accurate selection of wavelength regions of interest for single-point sensing applications and the integration of large, complex molecules for which line-based simulation is not available into datasets for machine learning applications. With a focus on photoacoustic spectral measurements, the generation of photoacoustic synthetic spectral data was presented. The synthetic data generation method allows to integrate known sensor parameters and the non-radiative relaxation effects. Such synthetic photoacoustic data can be used when a complex spectral system, including non-linear effects, is under investigation. While simple linear methods trained directly on measurement data remain the best choice for linear spectral applications, synthetic data can be applied when non-linearities limit the application of such simpler machine learning methods. As such, these methods can enable more complex sensor development to drive new applications and research frontiers.

# References

[1] J. Hodgkinson and R. P. Tatam, "Optical gas sensing: A review," *Measurement Science and Technology*, vol. 24, no. 1, 2013, ISSN: 0957-0233. DOI: 10.1088/0957-0233/24/1/012004.

[2] M. Phillips, "Breath tests in medicine," *Scientific American*, vol. 267, no. 1, pp. 74–79, 1992, ISSN: 0036-8733. DOI: 10.1038/scientificamerican0792-74.

[3] A. D. Subali, L. Wiyono, M. Yusuf, and M. F. A. Zaky, "The potential of volatile organic compounds-based breath analysis for covid-19 screening: A systematic review & meta-analysis," *Diagnostic microbiology and infectious disease*, vol. 102, no. 2, 2022. DOI: 10.1016/j.diagmicrobio.2021.115589.

[4] J. Pereira, P. Porto-Figueira, C. Cavaco, *et al.*, "Breath analysis as a potential and non-invasive frontier in disease diagnosis: An overview," *Metabolites*, vol. 5, no. 1, pp. 3–55, 2015, ISSN: 2218-1989. DOI: 10.3390/metabo5010003.

[5] D. C. Dumitras, M. Petrus, A.-M. Bratu, and C. Popa, "Applications of near infrared photoacoustic spectroscopy for analysis of human respiration: A review," *Molecules*, vol. 25, no. 7, 2020. DOI: 10.3390/molecules25071728.

[6] A. Amann, W. Miekisch, J. Schubert, *et al.*, "Analysis of exhaled breath for disease detection," *Annual review of analytical chemistry (Palo Alto, Calif.)*, vol. 7, pp. 455–482, 2014. DOI: 10.1146/annurev-anchem-071213-020043.

[7] G. C. Toon, C. B. Farmer, P. W. Schaper, L. L. Lowes, and R. H. Norton, "Composition measurements of the 1989 arctic winter stratosphere by airborne infrared solar absorption spectroscopy," *Journal of Geophysical Research*, vol. 97, no. D8, p. 7939, 1992, ISSN: 0148-0227. DOI: 10.1029/91JD03114.

[8] S. Palzer, "Photoacoustic-based gas sensing: A review," *Sensors (Basel, Switzerland)*, vol. 20, no. 9, 2020. DOI: 10.3390/s20092745.

[9] A. Sampaolo, P. Patimisco, M. Giglio, *et al.*, "Quartz-enhanced photoacoustic spectroscopy for multi-gas detection: A review," *Analytica chimica acta*, vol. 1202, 2022. DOI: 10.1016/j.aca.2021.338894.

[10] B. Lang, P. Breitegger, G. Brunnhofer, *et al.*, "Molecular relaxation effects on vibrational water vapor photoacoustic spectroscopy in air," *Applied Physics B*, vol. 126, no. 4, 2020, ISSN: 0946-2171. DOI: 10.1007/s00340-020-7409-3.

[11] P. Mishra, D. Passos, F. Marini, *et al.*, "Deep learning for near-infrared spectral data modelling: Hypes and benefits," *TrAC Trends in Analytical Chemistry*, vol. 157, 2022, ISSN: 01659936. DOI: 10.1016/j.trac.2022.116804.

[12] J. Goldschmidt, L. Nitzsche, S. Wolf, A. Lambrecht, and J. Wöllenstein, "Rapid quantitative analysis of ir absorption spectra for trace gas detection by artificial neural networks trained with synthetic data," *Sensors*, vol. 22, no. 3, p. 857, 2022. DOI: 10.3390/s22030857.

[13] V. V. Prischepa, V. E. Skiba, D. A. Vrazhnov, and Y. Kistenev, "Gas mixtures ir absorption spectra decomposition using a deep neural network," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 301, 2023, ISSN: 00224073. DOI: 10.1016/j.jqsrt.2023.108521.

[14] I. E. Gordon, L. S. Rothman, R. J. Hargreaves, *et al.*, "The hitran2020 molecular spectroscopic database," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 277, 2022. DOI: 10.1016/j.jqsrt.2021.107949.

[15] G. C. Toon, J.-F. Blavier, B. Sen, *et al.*, "Comparison of mkiv balloon and er-2 aircraft measurements of atmospheric trace gases," *Journal of Geophysical Research*, vol. 104, no. D21, pp. 26 779–26 790, 1999, ISSN: 0148-0227. DOI: 10.1029/1999JD900379.

[16] M. Müller, T. Rück, S. Jobst, *et al.*, "An algorithmic approach to compute the effect of non-radiative relaxation processes in photoacoustic spectroscopy," *Photoacoustics*, vol. 26, 2022, ISSN: 22135979. DOI: 10.1016/j.pacs.2022.100371.

[17] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Cham: Springer International Publishing, 2021, vol. 174, ISBN: 978-3-030-75177-7. DOI: 10.1007/978-3-030-75178-4.

[18] P. W. Atkins and C. A. Trapp, *Physikalische Chemie*, 3., korrigierte Aufl. Weinheim: Wiley-VCH, 2001, ISBN: 3-527-30236-0.

[19] S. Weigl, "Development of a sensor system for human breath acetone analysis based on photoacoustic spectroscopy," Ph. D. thesis, University of Regensburg, Regensburg, 2020.

[20] M. Hesse, H. Meier, B. Zeeh, S. Bienz, L. Bigler, and T. Fox, *Spektroskopische Methoden in der organischen Chemie*, 8., überarbeitete und erweiterte Auflage. Stuttgart and New York: Georg Thieme Verlag, 2012, ISBN: 9783135761084.

[21] International Organization for Standardization, *Iso 20473:2007 optics and photonics - spectral bands*, 2007.

[22] G. Wiegleb, *Gasmesstechnik in Theorie und Praxis: Messgeräte, Sensoren, Anwendungen*. Wiesbaden: Springer Vieweg, 2016, ISBN: 978-3-658-10686-7.

[23] J. Fraunhofer, "Bestimmung des brechungs-und des farbenzerstreungs-vermögens verschiedener glasarten, in bezug auf die vervollkommnung achromatischer fernröhre," *Annalen der Physik*, vol. 56, no. 5, pp. 264–313, 1817. DOI: 10.1002/andp.18170560706.

[24] E. Schrödinger, "An undulatory theory of the mechanics of atoms and molecules," *Physical Review*, vol. 28, no. 6, pp. 1049–1070, 1926, ISSN: 0031-899X. DOI: 10.1103/PhysRev.28.1049.

[25] P. M. Morse, "Diatomic molecules according to the wave mechanics. ii. vibrational levels," *Physical Review*, vol. 34, no. 1, pp. 57–64, 1929, ISSN: 0031-899X. DOI: 10.1103/PhysRev.34.57.

[26] W. Voigt, *Über das gesetz der intensitätsverteilung innerhalb der linien eines gasspektrums*, München, 1912.

[27] R. V. Kochanov, I. E. Gordon, L. S. Rothman, P. Wcisło, C. Hill, and J. S. Wilzewski, "Hitran application programming interface (hapi): A comprehensive approach to working with spectroscopic data," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 177, pp. 15–30, 2016, ISSN: 00224073. DOI: 10.1016/j.jqsrt.2016.03.005.

[28] S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick, and P. A. Johnson, "Gas-phase databases for quantitative infrared spectroscopy," *Applied Spectroscopy*, vol. 58, pp. 1452–1461, 2004. DOI: 10.1366/0003702042641281.

[29] L. S. Rothman, C. P. Rinsland, A. Goldman, *et al.*, "The hitran molecular spectroscopic database and hawks (hitran atmospheric workstation): 1996 edition," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 60, no. 5, pp. 665–710, 1998. DOI: 10.1016/S0022-4073(98)00078-8.

[30] R. R. Gamache, B. Vispoel, M. Rey, *et al.*, "Total internal partition sums for the hitran2020 database," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 271, 2021. DOI: 10.1016/j.jqsrt.2021.107713.

[31] J.-M. Hartmann, C. Boulet, and D. Robert, *Collisional Effects on Molecular Spectra*. Elsevier, 2008, ISBN: 9780444520173. DOI: 10.1016/B978-0-444-52017-3.X0001-5.

[32] W. Heisenberg, "Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik," *Zeitschrift für Physik*, vol. 43, no. 3-4, pp. 172–198, 1927, ISSN: 0044-3328. DOI: 10.1007/BF01397280.

[33] C. Doppler, "Über das farbige licht der doppelsterne und einiger anderer gestirne des himmels: Versuch einer das bradley'sche aberrations-theorem als integrirenden theil in sich schliessenden allgemeineren theorie," vol. V, no. 2, pp. 465–482, 1842.

[34] R. Brown, "Xxvii. a brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *The Philosophical Magazine*, vol. 4, no. 21, pp. 161–173, 1828, ISSN: 1941-5850. DOI: 10.1080/14786442808674769.

[35] A. Einstein, "Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen," *Annalen der Physik und Chemie*, vol. 322, no. 8, pp. 549–560, 1905, ISSN: 00033804. DOI: 10.1002/andp.19053220806.

[36] J. C. Maxwell, "Ii. illustrations of the dynamical theory of gases," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 20, no. 130, pp. 21–37, 1860, ISSN: 1941-5982. DOI: 10.1080/14786446008642902.

[37] J. C. Maxwell, "V. illustrations of the dynamical theory of gases. —part i. on the motions and collisions of perfectly elastic spheres," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 19, no. 124, pp. 19–32, 1860, ISSN: 1941-5982. DOI: 10.1080/14786446008642818.

[38] H. A. Lorentz, "The absorption and emission lines of gaseous bodies," in *Knaw, Proceedings*, vol. 8, 1906, pp. 591–611. (visited on 11/29/2021).

[39] H. Margenau, "Pressure shift and broadening of spectral lines," *Physical Review*, vol. 40, no. 3, pp. 387–408, 1932, ISSN: 0031-899X. DOI: 10.1103/PhysRev.40.387.

[40] S. G. Rautian and I. I. Sobel'man, "The effect of collisions on the doppler broadening of spectral lines," *Soviet Physics Uspekhi*, vol. 9, no. 5, pp. 701–716, 1967, ISSN: 0038-5670. DOI: 10.1070/PU1967v009n05ABEH003212.

[41] J. P. Wittke and R. H. Dicke, "Redetermination of the hyperfine splitting in the ground state of atomic hydrogen," *Physical Review*, vol. 103, no. 3, p. 620, 1956, ISSN: 0031-899X. DOI: 10.1103/PhysRev.103.6209.

[42] R. Bogue, "Detecting gases with light: A review of optical gas sensor technologies," *Sensor Review*, vol. 35, no. 2, pp. 133–140, 2015, ISSN: 0260-2288. DOI: 10.1108/SR-09-2014-696.

[43] J. H. Lambert, *Photometria, sive De mensura et gradibus luminis, colorum et umbrae*. Augsburg: Sumptibus Vidvae Eberhardi Klett, 1760.

[44] A. Beer, "Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten," *Annalen der Physik und Chemie*, vol. 162, no. 5, pp. 78–88, 1852, ISSN: 00033804. DOI: 10.1002/andp.18521620505.

[45] K. F. Luft, "Über eine neue methode der registrierenden gasanalyse mit hilfe der absorption ultraroter strahlen ohne spektrale zerlegung," *Z. tech. Phys*, vol. 24, pp. 97–104, 1943.

[46] A. A. Michelson, "Xxviii. interference phenomena in a new form of refractometer," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 81, pp. 236–242, 1882, ISSN: 1941-5982. DOI: 10.1080/14786448208627176.

[47] A. G. Bell, "On the production and reproduction of sound by light," *American Journal of Science*, vol. s3-20, no. 118, pp. 305–324, 1880, ISSN: 0002-9599. DOI: 10.2475/ajs.s3-20.118.305.

[48] M. L. Viegerov, "Eine methode der gasanalyse, beruhend auf der optisch-akustischen tyndall-röntgenerscheinung," in *Dolady Akademii Nauk SSSR*, vol. 19, pp. 687–688.

[49] S. Weigl, M. Müller, J. Pangerl, and T. Rück, "Scopes and limits of photoacoustic spectroscopy in modern breath analysis," in *Breath Analysis*, ser. Bioanalytical Reviews, S. Weigl, Ed., vol. 4, Cham: Springer International Publishing, 2023, pp. 101–159, ISBN: 978-3-031-18525-0. DOI: 10.1007/11663-2022-22.

[50] P. Hess, "Resonant photoacoustic spectroscopy," in *Physical and Inorganic Chemistry*, ser. Topics in Current Chemistry, F. L. Boschke, M. J. S. Dewar, J. D. Dunitz, *et al.*, Eds., vol. 111, Berlin, Heidelberg: Springer Berlin Heidelberg, 1983, pp. 1–32, ISBN: 978-3-540-12065-0. DOI: 10.1007/3-540-12065-3-1.

[51] A. Miklós, S. Schäfer, and P. Hess, "Photoacoustic spectroscopy, theory," in *Encyclopedia of Spectroscopy and Spectrometry*, Elsevier, 1999, pp. 1815–1822, ISBN: 9780122266805. DOI: 10.1006/rwsp.2000.0234.

[52] K. Liu, H. Yi, A. A. Kosterev, *et al.*, "Trace gas detection based on off-beam quartz enhanced photoacoustic spectroscopy: Optimization and performance evaluation," *The Review of scientific instruments*, vol. 81, no. 10, 2010. DOI: 10.1063/1.3480553.

[53] J. Pangerl, M. Müller, T. Rück, S. Weigl, and R. Bierl, "Characterizing a sensitive compact mid-infrared photoacoustic sensor for methane, ethane and acetylene detection considering changing ambient parameters and bulk composition (n2, o2 and h2o)," *Sensors and Actuators B: Chemical*, vol. 352, 2022, ISSN: 09254005. DOI: 10.1016/j.snb.2021.130962.

[54] D. A. Russell, "On the sound field radiated by a tuning fork," *American Journal of Physics*, vol. 68, no. 12, pp. 1139–1145, 2000, ISSN: 0002-9505. DOI: 10.1119/1.1286661.

[55] T. Rück, R. Bierl, and F.-M. Matysik, "No2 trace gas monitoring in air using off-beam quartz enhanced photoacoustic spectroscopy (qepas) and interference studies towards co2, h2o and acoustic noise," *Sensors and Actuators B: Chemical*, vol. 255, pp. 2462–2471, 2018, ISSN: 09254005. DOI: 10.1016/j.snb.2017.09.039.

[56] L. Liu, H. Huan, A. Mandelis, *et al.*, "Design and structural optimization of t-resonators for highly sensitive photoacoustic trace gas detection," *Optics & Laser Technology*, vol. 148, 2022, ISSN: 00303992. DOI: 10.1016/j.optlastec.2021.107695.

[57] D. Hofstetter, M. Beck, J. Faist, M. Nägele, and M. W. Sigrist, "Photoacoustic spectroscopy with quantum cascade distributed-feedback lasers," *Optics letters*, vol. 26, no. 12, pp. 887–889, 2001, ISSN: 0146-9592. DOI: 10.1364/OL.26.000887.

[58] F. Ma, Z. Liao, Y. Zhao, *et al.*, "Detection of trace c2h2 in n2 buffer gas with cantilever-enhanced photoacoustic spectrometer," *Optik*, vol. 232, 2021, ISSN: 00304026. DOI: 10.1016/j.ijleo.2021.166525.

[59] L. Dong, R. Lewicki, K. Liu, P. R. Buerki, M. J. Weida, and F. K. Tittel, "Ultra-sensitive carbon monoxide detection by using ec-qcl based quartz-enhanced photoacoustic spectroscopy," *Applied Physics B*, vol. 107, no. 2, pp. 275–283, 2012, ISSN: 0946-2171. DOI: 10.1007/s00340-012-4949-1.

[60] F. Sgobba, A. Sampaolo, P. Patimisco, *et al.*, "Compact and portable quartz-enhanced photoacoustic spectroscopy sensor for carbon monoxide environmental monitoring in urban areas," *Photoacoustics*, vol. 25, 2022, ISSN: 22135979. DOI: 10.1016/j.pacs.2021.100318.

[61] J. Hayden, B. Baumgartner, and B. Lendl, "Anomalous humidity dependence in photoacoustic spectroscopy of co explained by kinetic cooling," *Applied Sciences*, vol. 10, no. 3, p. 843, 2020. DOI: 10.3390/app10030843.

[62] S. Weigl, E. Wittmann, T. Rück, R. Bierl, and F.-M. Matysik, "Effects of ambient parameters and cross-sensitivities from o2, co2 and h2o on the photoacoustic detection of acetone in the uv region," *Sensors and Actuators B: Chemical*, vol. 328, 2021, ISSN: 09254005. DOI: 10.1016/j.snb.2020.129001.

[63] A. J. Zuckerwar, *Handbook of the Speed of Sound in Real Gases*. San Diego, CA: Elsevier, 2002.

[64] F. Herbert, “Spectrum line profiles: A generalized voigt function including collisional narrowing,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 14, no. 9, pp. 943–951, 1974. DOI: 10.1016/0022-4073(74)90021-1.

[65] P. L. Varghese and R. K. Hanson, “Collisional narrowing effects on spectral line shapes measured at high resolution,” *Applied optics*, vol. 23, no. 14, p. 2376, 1984, ISSN: 1559-128X. DOI: 10.1364/AO.23.002376.

[66] B. H. Armstrong, “Spectrum line profiles: The voigt function,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 7, no. 1, pp. 61–88, 1967. DOI: 10.1016/0022-4073(67)90057-X.

[67] M. Abramowitz and I. A. Stegun, “Handbook of mathematical functions with formulas, graphs, and mathematical tables. national bureau of standards applied mathematics series 55.,” 1972.

[68] J. Humlíček, “Optimized computation of the voigt and complex probability functions,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 27, no. 4, pp. 437–444, 1982. DOI: 10.1016/0022-4073(82)90078-4.

[69] J. A. C. Weideman, “Computation of the complex error function,” *SIAM Journal on Numerical Analysis*, vol. 31, no. 5, pp. 1497–1518, 1994, ISSN: 0036-1429. DOI: 10.1137/0731077.

[70] F. Schreier, “Optimized implementations of rational approximations for the voigt and complex error function,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 112, no. 6, pp. 1010–1025, 2011. DOI: 10.1016/j.jqsrt.2010.12.010.

[71] J. Tennyson, P. F. Bernath, A. Campargue, *et al.*, “Recommended isolated-line profile for representing high-resolution spectroscopic transitions (iupac technical report),” *Pure and Applied Chemistry*, vol. 86, no. 12, pp. 1931–1943, 2014, ISSN: 0033-4545. DOI: 10.1515/pac-2014-0208.

[72] N. H. Ngo, D. Lisak, H. Tran, and J.-M. Hartmann, “An isolated line-shape model to go beyond the voigt profile in spectroscopic databases and radiative transfer codes,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 129, pp. 89–100, 2013. DOI: 10.1016/j.jqsrt.2013.05.034.

[73] F. Rohart, H. Mäder, and H.-.-W. Nicolaisen, “Speed dependence of rotational relaxation induced by foreign gas collisions: Studies on ch 3 f by millimeter wave coherent transients,” *The Journal of Chemical Physics*, vol. 101, no. 8, pp. 6475–6486, 1994, ISSN: 0021-9606. DOI: 10.1063/1.468342.

[74] D. Albert, B. K. Antony, Y. A. Ba, *et al.*, “A decade with vamdc: Results and ambitions,” *Atoms*, vol. 8, no. 4, p. 76, 2020. DOI: 10.3390/atoms8040076.

[75] M. L. Dubernet, V. Boudon, J. L. Culhane, *et al.*, “Virtual atomic and molecular data centre,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 111, no. 15, pp. 2151–2159, 2010. DOI: 10.1016/j.jqsrt.2010.05.004.

[76] P. F. Bernath, "Mollist: Molecular line lists, intensities and spectra," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 240, 2020. DOI: 10.1016/j.jqsrt.2019.106687.

[77] I. E. Gordon, L. S. Rothman, C. Hill, *et al.*, "The hitran2016 molecular spectroscopic database," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 203, pp. 3–69, 2017. DOI: 10.1016/j.jqsrt.2017.06.038.

[78] R.V. Kochanov, I.E. Gordon, L.S. Rothman, K.P. Shine, S.W. Sharpe, T.J. Johnson, T.J. Wallington, J.J. Harrison, P.F. Bernath, M. Birk, G. Wagner, K. Le Bris, I. Bravo, C. Hill, "Infrared absorption cross-sections in hitran2016 and beyond: Expansion for climate, environment, and atmospheric applications," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 230, pp. 172–221, 2019. DOI: 10.1016/j.jqsrt.2019.04.001.

[79] J. Ballard, W. B. Johnston, M. R. Gunson, and P. T. Wassell, "Absolute absorption coefficients of clono 2 infrared bands at stratospheric temperatures," *Journal of Geophysical Research*, vol. 93, no. D2, p. 1659, 1988, ISSN: 0148-0227. DOI: 10.1029/JD093iD02p01659.

[80] J. J. Michalenko, C. M. Murzyn, J. D. Zollweg, L. Wermer, A. J. van Omen, and M. D. Clemenson, "Machine learning predictions of transition probabilities in atomic spectra," *Atoms*, vol. 9, no. 1, p. 2, 2021. DOI: 10.3390/atoms9010002.

[81] Jet Propulsion Laboratory, *Pseudo linelists*. [Online]. Available: https://mark4sun.jpl.nasa.gov/pseudo.html (visited on 09/29/2021).

[82] K. Sung, G. C. Toon, A. W. Mantz, and M. A. H. Smith, "Ft-ir measurements of cold c3h8 cross sections at 7–15µm for titan atmosphere," *Icarus*, vol. 226, no. 2, pp. 1499–1513, 2013, ISSN: 00191035. DOI: 10.1016/j.icarus.2013.07.028.

[83] K. Sung, B. Steffens, G. C. Toon, D. J. Nemchick, and M. A. H. Smith, "Pseudoline parameters to represent n-butane (n-c4h10) cross-sections measured in the 7–15 µm region for the titan atmosphere," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 251, 2020. DOI: 10.1016/j.jqsrt.2020.107011.

[84] B. R. Kowalski, "Chemometrics," *Analytical Chemistry*, vol. 52, no. 5, pp. 112–122, 1980, ISSN: 0003-2700. DOI: 10.1021/ac50055a016.

[85] S. Guo, J. Popp, and T. Bocklitz, "Chemometric analysis in raman spectroscopy from experimental design to machine learning-based modeling," *Nature protocols*, vol. 16, no. 12, pp. 5426–5459, 2021. DOI: 10.1038/s41596-021-00620-3.

[86] S. Kern, S. Liehr, L. Wander, *et al.*, "Artificial neural networks for quantitative online nmr spectroscopy," *Analytical and bioanalytical chemistry*, vol. 412, no. 18, pp. 4447–4459, 2020. DOI: 10.1007/s00216-020-02687-5.

[87] R. Houhou and T. Bocklitz, "Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data," *Analytical Science Advances*, vol. 2, no. 3-4, pp. 128–141, 2021, ISSN: 2628-5452. DOI: 10.1002/ansa.202000162.

[88] W. C. Mundy, J. A. Roux, and A. M. Smith, "Mie scattering by spheres in an absorbing medium," *Journal of the Optical Society of America*, vol. 64, no. 12, p. 1593, 1974, ISSN: 0030-3941. DOI: 10.1364/JOSA.64.001593.

[89] R. G. Brereton, "Pattern recognition in chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 90–96, 2015, ISSN: 01697439. DOI: 10.1016/j.chemolab.2015.06.012.

[90] P. Geladi and K. Esbensen, "The start and early history of chemometrics: Selected interviews. part 1," *Journal of Chemometrics*, vol. 4, no. 5, pp. 337–354, 1990, ISSN: 0886-9383. DOI: 10.1002/cem.1180040503.

[91] K. Esbensen and P. Geladi, "The start and early history of chemometrics: Selected interviews. part 2," *Journal of Chemometrics*, vol. 4, no. 6, pp. 389–412, 1990, ISSN: 0886-9383. DOI: 10.1002/cem.1180040604.

[92] J. J. Workman, P. R. Mobley, B. R. Kowalski, and R. Bro, "Review of chemometrics applied to spectroscopy: 1985-95, part i," *Applied Spectroscopy Reviews*, vol. 31, no. 1-2, pp. 73–124, 1996, ISSN: 0570-4928. DOI: 10.1080/05704929608000565.

[93] P. R. Mobley, B. R. Kowalski, J. J. Workman, and R. Bro, "Review of chemometrics applied to spectroscopy: 1985-95, part 2," *Applied Spectroscopy Reviews*, vol. 31, no. 4, pp. 347–368, 1996, ISSN: 0570-4928. DOI: 10.1080/05704929608000575.

[94] R. Bro, J. J. Workman, P. R. Mobley, and B. R. Kowalski, "Review of chemometrics applied to spectroscopy: 1985-95, part 3 — multi-way analysis," *Applied Spectroscopy Reviews*, vol. 32, no. 3, pp. 237–261, 1997, ISSN: 0570-4928. DOI: 10.1080/05704929708003315.

[95] R. G. Brereton, J. Jansen, J. Lopes, *et al.*, "Chemometrics in analytical chemistry-part i: History, experimental design and data analysis tools," *Analytical and bioanalytical chemistry*, vol. 409, no. 25, pp. 5891–5899, 2017. DOI: 10.1007/s00216-017-0517-1.

[96] R. G. Brereton, J. Jansen, J. Lopes, *et al.*, "Chemometrics in analytical chemistry-part ii: Modeling, validation, and applications," *Analytical and bioanalytical chemistry*, vol. 410, no. 26, pp. 6691–6704, 2018. DOI: 10.1007/s00216-018-1283-4.

[97] C. A. Meza Ramirez, M. Greenop, L. Ashton, and I. u. Rehman, "Applications of machine learning in spectroscopy," *Applied Spectroscopy Reviews*, vol. 56, no. 8-10, pp. 733–763, 2021, ISSN: 0570-4928. DOI: 10.1080/05704928.2020.1859525.

[98] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. C. Buydens, and E. Marchiori, "Convolutional neural networks for vibrational spectroscopic data analysis," *Analytica chimica acta*, vol. 954, pp. 22–31, 2017. DOI: 10.1016/j.aca.2016.12.010.

[99] X. Zhang, T. Lin, J. Xu, X. Luo, and Y. Ying, "Deepspectra: An end-to-end deep learning approach for quantitative spectral analysis," *Analytica chimica acta*, vol. 1058, pp. 48–57, 2019. DOI: 10.1016/j.aca.2019.01.002.

[100] P. Fu, Y. Wen, Y. Zhang, *et al.*, "Spectratr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure," *Journal of Innovative Optical Health Sciences*, vol. 15, no. 03, 2022, ISSN: 1793-5458. DOI: 10.1142/S1793545822500213.

[101] C. M. Bishop, *Pattern recognition and machine learning* (Information science and statistics), Corrected at 8th printing 2009. New York, NY: Springer, 2009, ISBN: 978-1-4939-3843-8.

[102] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (Adaptive computation and machine learning). Cambridge, Massachusetts and London, England: The MIT Press, 2016, ISBN: 9780262035613.

[103] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach* (Prentice-Hall series in artificial intelligence), 3. ed. Upper Saddle River, NJ: Prentice-Hall, 2010, ISBN: 0136042597.

[104] P. Wang, "On defining artificial intelligence," *Journal of Artificial General Intelligence*, vol. 10, no. 2, pp. 1–37, 2019. DOI: 10.2478/jagi-2019-0002.

[105] A. de Juan, J. Jaumot, and R. Tauler, "Multivariate curve resolution (mcr). solving the mixture analysis problem," *Anal. Methods*, vol. 6, no. 14, pp. 4964–4976, 2014, ISSN: 1759-9660. DOI: 10.1039/C4AY00571F.

[106] A. de Juan and R. Tauler, "Multivariate curve resolution: 50 years addressing the mixture analysis problem - a review," *Analytica chimica acta*, vol. 1145, pp. 59–78, 2021. DOI: 10.1016/j.aca.2020.10.051.

[107] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001, ISSN: 01697439. DOI: 10.1016/S0169-7439(01)00155-1.

[108] K. Kumar, "Partial least square (pls) analysis," *Resonance*, vol. 26, no. 3, pp. 429–442, 2021, ISSN: 0971-8044. DOI: 10.1007/s12045-021-1140-1.

[109] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction* (Springer series in statistics), 2. ed., corr. at 4. print. New York, NY: Springer, 2009, ISBN: 978-0-387-84858-7.

[110] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. DOI: 10.1145/3065386.

[111] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2016. DOI: 10.1109/CVPR.2016.90.

[112] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[113] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, L. Rokach, O. Maimon, and E. Shmueli, Eds. Cham: Springer International Publishing, 2023, pp. 353–374, ISBN: 978-3-031-24628-9. DOI: 10.1007/978-3-031-24628-9_16.

[114] P. Mishra and D. Passos, "Deep calibration transfer: Transferring deep learning models between infrared spectroscopy instruments," *Infrared Physics & Technology*, vol. 117, 2021, ISSN: 1350-4495. DOI: 10.1016/j.infrared.2021.103863.

[115] P. Mishra and D. Passos, "Realizing transfer learning for updating deep learning models of spectral data to be used in a new scenario," *Chemometrics and Intelligent Laboratory Systems*, 2021, ISSN: 01697439. DOI: 10.1016/j.chemolab.2021.104283.

[116] R. Lewicki, G. Wysocki, A. A. Kosterev, and F. K. Tittel, "Qepas based detection of broadband absorbing molecules using a widely tunable, cw quantum cascade laser at 8.4 mum," *Optics express*, vol. 15, no. 12, pp. 7357–7366, 2007. DOI: 10.1364/OE.15.007357.

[117] A. A. Kosterev, P. R. Buerki, L. Dong, M. Reed, T. Day, and F. K. Tittel, "Qepas detector for rapid spectral measurements," *Applied Physics B*, vol. 100, no. 1, pp. 173–180, 2010, ISSN: 0946-2171. DOI: 10.1007/s00340-010-3975-0.

[118] Y. V. Kistenev, A. V. Borisov, D. A. Kuzmin, *et al.*, "Breath air measurement using wide-band frequency tuning ir laser photo-acoustic spectroscopy," in *Dynamics and Fluctuations in Biomedical Photonics XIII*, V. V. Tuchin, K. V. Larin, M. J. Leahy, and R. K. Wang, Eds., ser. SPIE Proceedings, SPIE, 2016. DOI: 10.1117/12.2214645.

[119] Y. Kistenev, A. Borisov, V. Nikolaev, D. Vrazhnov, and D. Kuzmin, "Laser photoacoustic spectroscopy applications in breathomics," *Journal of Biomedical Photonics & Engineering*, vol. 5, no. 1, 2019. DOI: 10.18287/JBPE19.05.010303.

[120] A. V. Borisov, A. G. Syrkina, D. A. Kuzmin, *et al.*, "Application of machine learning and laser optical-acoustic spectroscopy to study the profile of exhaled air volatile markers of acute myocardial infarction," *Journal of breath research*, vol. 15, no. 2, 2021. DOI: 10.1088/1752-7163/abebd4.

[121] G. Menduni, F. Sgobba, S. D. Russo, *et al.*, "Fiber-coupled quartz-enhanced photoacoustic spectroscopy system for methane and ethane monitoring in the near-infrared spectral range," *Molecules (Basel, Switzerland)*, vol. 25, no. 23, 2020. DOI: 10.3390/molecules25235607.

[122] G. Menduni, A. Zifarelli, A. Sampaolo, *et al.*, "High-concentration methane and ethane qepas detection employing partial least squares regression to filter out energy relaxation dependence on gas matrix composition," *Photoacoustics*, vol. 26, 2022, ISSN: 2213-5979. DOI: 10.1016/j.pacs.2022.100349.

[123] M. Giglio, A. Zifarelli, A. Sampaolo, *et al.*, "Broadband detection of methane and nitrous oxide using a distributed-feedback quantum cascade laser array and quartz-enhanced photoacoustic sensing," *Photoacoustics*, vol. 17, 2020, ISSN: 2213-5979. DOI: 10.1016/j.pacs.2019.100159.

[124] A. Zifarelli, M. Giglio, G. Menduni, *et al.*, "Partial least-squares regression as a tool to retrieve gas concentrations in mixtures detected using quartz-enhanced photoacoustic spectroscopy," *Analytical Chemistry*, vol. 92, no. 16, pp. 11 035–11 043, 2020, ISSN: 0003-2700. DOI: 10.1021/acs.analchem.0c00075.

[125] J. Pangerl, E. Moser, M. Müller, *et al.*, "A sub-ppbv-level acetone and ethanol quantum cascade laser based photoacoustic sensor - characterization and multi-component spectra recording in synthetic breath," *Photoacoustics*, vol. 30, 2023, ISSN: 2213-5979. DOI: 10.1016/j.pacs.2023.100473.

[126] A. A. Karapuzikov, I. V. Sherstov, D. B. Kolker, *et al.*, "Laserbreeze gas analyzer for noninvasive diagnostics of air exhaled by patients," *Physics of Wave Phenomena*, vol. 22, no. 3, pp. 189–196, 2014, ISSN: 1541-308X. DOI: 10.3103/S1541308X14030054.

[127] F. Vernuccio, A. Bresci, V. Cimini, *et al.*, "Artificial intelligence in classical and quantum photonics," *Laser & Photonics Reviews*, vol. 16, no. 5, 2022, ISSN: 1863-8880. DOI: 10.1002/lpor.202100399.

[128] N. T. Anderson and K. B. Walsh, "Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation," *Journal of Near Infrared Spectroscopy*, vol. 30, no. 1, pp. 3–17, 2022, ISSN: 0967-0335. DOI: 10.1177/09670335211057235.

[129] J. Yang, J. Xu, X. Zhang, C. Wu, T. Lin, and Y. Ying, "Deep learning for vibrational spectral analysis: Recent progress and a practical guide," *Analytica chimica acta*, vol. 1081, pp. 6–17, 2019. DOI: 10.1016/j.aca.2019.06.012.

[130] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Analytica chimica acta*, vol. 648, no. 1, pp. 77–84, 2009. DOI: 10.1016/j.aca.2009.06.046.

[131] B. Nagy, D. L. Galata, A. Farkas, and Z. K. Nagy, "Application of artificial neural networks in the process analytical technology of pharmaceutical manufacturing-a review," *The AAPS journal*, vol. 24, no. 4, p. 74, 2022. DOI: 10.1208/s12248-022-00706-0.

[132] A. A. Boateng, S. Sumaila, M. Lartey, M. B. Oppong, K. F. Opuni, and L. A. Adutwum, "Evaluation of chemometric classification and regression models for the detection of syrup adulteration in honey," *LWT*, vol. 163, 2022, ISSN: 00236438. DOI: 10.1016/j.lwt.2022.113498.

[133] T. Voumard, T. Wildi, V. Brasch, R. G. Álvarez, G. V. Ogando, and T. Herr, "Ai-enabled real-time dual-comb molecular fingerprint imaging," *Optics letters*, vol. 45, no. 24, pp. 6583–6586, 2020. DOI: 10.1364/OL.410762.

[134] C. Cui and T. Fearn, "Modern practical convolutional neural networks for multivariate regression: Applications to nir calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 182, pp. 9–20, 2018, ISSN: 01697439. DOI: 10.1016/j.chemolab.2018.07.008.

[135] J. Padarian, B. Minasny, and A. B. McBratney, "Using deep learning to predict soil properties from regional spectral data," *Geoderma Regional*, vol. 16, 2019, ISSN: 23520094. DOI: 10.1016/j.geodrs.2018.e00198.

[136] L. B. Ayres, F. J. V. Gomez, J. R. Linton, M. F. Silva, and C. D. Garcia, "Taking the leap between analytical chemistry and artificial intelligence: A tutorial review," *Analytica chimica acta*, vol. 1161, 2021. DOI: 10.1016/j.aca.2021.338403.

[137] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for raman and infrared spectroscopy," *EPJ Techniques and Instrumentation*, vol. 2, no. 1, 2015. DOI: 10.1140/epjti/s40485-015-0018-6.

[138] Å. Rinnan, F. den van Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009, ISSN: 01659936. DOI: 10.1016/j.trac.2009.07.007.

[139] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 92–99, 2012, ISSN: 01697439. DOI: 10.1016/j.chemolab.2012.03.004.

[140] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964, ISSN: 0003-2700. DOI: 10.1021/ac60214a047.

[141] E. J. Bjerrum, M. Glahder, and T. Skov, *Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics*, 2017. DOI: 10.48550/arXiv.1710.01927.

[142] P. Mishra and D. Passos, "A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit," *Chemometrics and Intelligent Laboratory Systems*, vol. 212, 2021, ISSN: 01697439. DOI: 10.1016/j.chemolab.2021.104287.

[143] P. Mishra, J. M. Roger, D. N. Rutledge, and E. Woltering, "Sport pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials," *Postharvest Biology and Technology*, vol. 168, 2020, ISSN: 09255214. DOI: 10.1016/j.postharvbio.2020.111271.

[144] E. A. Magnussen, B. Zimmermann, U. Blazhko, *et al.*, "Deep learning-enabled inference of 3d molecular absorption distribution of biological cells from ir spectra," *Communications Chemistry*, vol. 5, no. 1, 2022. DOI: 10.1038/s42004-022-00792-3.

[145] D. Passos and P. Mishra, "A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, 2022, ISSN: 01697439. DOI: 10.1016/j.chemolab.2022.104520.

[146] L. Biewald, *Experiment tracking with weights and biases*, Software available from wandb.com, 2020. [Online]. Available: https://www.wandb.com/.

[147] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[148] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012, ISSN: 01697439. DOI: 10.1016/j.chemolab.2012.07.010.

[149] T. Mehmood, S. Sæbø, and K. H. Liland, "Comparison of variable selection methods in partial least squares regression," *Journal of Chemometrics*, vol. 34, no. 6, 2020, ISSN: 0886-9383. DOI: 10.1002/cem.3226.

[150] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics (Oxford, England)*, vol. 23, no. 19, pp. 2507–2517, 2007. DOI: 10.1093/bioinformatics/btm344.

[151] A. G. Frenich, D. Jouan-Rimbaud, D. L. Massart, S. Kuttatharmmakul, M. M. Galera, and J. L. M. Vidal, "Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares," *The Analyst*, vol. 120, no. 12, p. 2787, 1995, ISSN: 0003-2654. DOI: 10.1039/AN9952002787.

[152] P. Mishra, E. Woltering, B. Brouwer, and E. Hogeveen-van Echtelt, "Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach," *Postharvest Biology and Technology*, vol. 171, 2021, ISSN: 09255214. DOI: 10.1016/j.postharvbio.2020.111348.

[153] J. M. Roger, B. Palagos, D. Bertrand, and E. Fernandez-Ahumada, "Covsel: Variable selection for highly multivariate and multi-response calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 2, pp. 216–223, 2011, ISSN: 01697439. DOI: 10.1016/j.chemolab.2010.10.003.

[154] J. Ghasemi, A. Niazi, and R. Leardi, "Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by pls: Application on copper and zinc mixture," *Talanta*, vol. 59, no. 2, pp. 311–317, 2003. DOI: 10.1016/S0039-9140(02)00505-2.

[155] B.-C. Deng, Y.-H. Yun, D.-S. Cao, *et al.*, "A bootstrapping soft shrinkage approach for variable selection in chemical modeling," *Analytica chimica acta*, vol. 908, pp. 63–74, 2016. DOI: 10.1016/j.aca.2016.01.001.

[156] Y.-H. Yun, W.-T. Wang, B.-C. Deng, *et al.*, "Using variable combination population analysis for variable selection in multivariate calibration," *Analytica chimica acta*, vol. 862, pp. 14–23, 2015. DOI: 10.1016/j.aca.2014.12.048.

[157] V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna, "Elimination of uninformative variables for multivariate calibration," *Analytical Chemistry*, vol. 68, no. 21, pp. 3851–3858, 1996, ISSN: 0003-2700. DOI: 10.1021/ac960321m.

[158] S. Sæbø, T. Almøy, J. Aarøe, and A. H. Aastveit, "St-pls: A multi-directional nearest shrunken centroid type classifier via pls," *Journal of Chemometrics*, vol. 22, no. 1, pp. 54–62, 2008, ISSN: 0886-9383. DOI: 10.1002/cem.1101.

[159] W. Kausch, S. Noll, A. Smette, *et al.*, "Molecfit: A general tool for telluric absorption correction," *Astronomy & Astrophysics*, vol. 576, A78, 2015, ISSN: 1432-0746. DOI: 10.1051/0004-6361/201423909.

[160] R. D. Kjærsgaard, A. Bello-Arufe, A. D. Rathcke, L. A. Buchhave, and L. K. H. Clemmensen, “Unsupervised spectral unmixing for telluric correction using a neural network autoencoder,” *Proceedings of Fourth Workshop on Machine Learning and the Physical Sciences*, 2021. DOI: arXiv:2111.09081v1.

[161] E. A. Magnussen, J. H. Solheim, U. Blazhko, *et al.*, “Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells,” *Journal of biophotonics*, vol. 13, no. 12, 2020. DOI: 10.1002/jbio.202000204.

[162] B.-X. Xue, M. Barbatti, and P. O. Dral, “Machine learning for absorption cross sections,” *The journal of physical chemistry. A*, vol. 124, no. 35, pp. 7199–7210, 2020. DOI: 10.1021/acs.jpca.0c05310.

[163] H. Henschel, A. T. Andersson, W. Jespers, M. Mehdi Ghahremanpour, and D. van der Spoel, “Theoretical infrared spectra: Quantitative similarity measures and force fields,” *Journal of chemical theory and computation*, vol. 16, no. 5, pp. 3307–3315, 2020. DOI: 10.1021/acs.jctc.0c00126.

[164] H. Henschel and D. van der Spoel, “An intuitively understandable quality measure for theoretical vibrational spectra,” *The journal of physical chemistry letters*, vol. 11, no. 14, pp. 5471–5475, 2020. DOI: 10.1021/acs.jpclett.0c01655.

[165] E. Moser, S. Jobst, R. Bierl, and F. Jenko, “A deep learning system to transform cross-section spectra to varying environmental conditions,” *Vibrational Spectroscopy*, 2022, ISSN: 09242031. DOI: 10.1016/j.vibspec.2022.103410.

[166] G. Wagner and M. Birk, “New infrared spectroscopic database for chlorine nitrate,” *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 82, no. 1-4, pp. 443–460, 2003, ISSN: 00224073. DOI: 10.1016/S0022-4073(03)00169-9.

[167] T. Schaar, “Disentangling latent spaces with (semi-) supervised adversarial autoencoders,” Master Thesis, Technical University of Munich, Munich, 2021.

[168] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, *Adversarial autoencoders*, Nov. 17, 2015. DOI: arXiv:1511.05644.

[169] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010. DOI: 10.1021/ci100050t.

[170] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, “Understanding batch normalization,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc, 2018.

[171] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, “How does batch normalization help optimization?” In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc, 2018.

[172] F. Hu, M. Zhou, P. Yan, *et al.*, "Identification of mine water inrush using laser-induced fluorescence spectroscopy combined with one-dimensional convolutional neural network," *RSC advances*, vol. 9, no. 14, pp. 7673–7679, 2019. DOI: 10.1039/C9RA00805E.

[173] N. Gudur, "Architectural study of effects of kernel sizes on chemometric data using convolutional neural networks," Master Thesis, OTH Regensburg, Regensburg, 2022.

[174] N. T. Anderson, K. B. Walsh, P. P. Subedi, and C. H. Hayes, "Achieving robustness across season, location and cultivar for a nirs model for intact mango fruit dry matter content," *Postharvest Biology and Technology*, vol. 168, 2020, ISSN: 09255214. DOI: 10.1016/j.postharvbio.2020.111202.

[175] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, and B. A. Moser, "Domain adaptation for regression under beer–lambert's law," *Knowledge-Based Systems*, vol. 210, 2020, ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106447.

[176] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods (San Diego, Calif.)*, vol. 71, pp. 58–63, 2015. DOI: 10.1016/j.ymeth.2014.08.005.

[177] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," *Journal of computer-aided molecular design*, vol. 30, no. 8, pp. 595–608, 2016. DOI: 10.1007/s10822-016-9938-8.

[178] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of mdl keys for use in drug discovery," *Journal of chemical information and computer sciences*, vol. 42, no. 6, pp. 1273–1280, 2002, ISSN: 0095-2338. DOI: 10.1021/ci010132r.

[179] J. Pangerl, E. Wittmann, S. Weigl, M. Müller, R. Bierl, and F.-M. Matysik, "Using a modulated quantum cascade laser for photoacoustic spectra recording of exhaled acetone and main breath components," in *Optical Sensors and Sensing Congress 2022 (AIS, LACSEA, Sensors, ES)*, Washington, D.C.: Optica Publishing Group, ISBN: 978-1-957171-10-4. DOI: 10.1364/AIS.2022.ATu3G.1.

[180] E. Moser, J. Pangerl, S. Jobst, S. Weigl, and R. Bierl, "Modeling the photoacoustic spectrum of a quantum cascade laser for human breath," in *Optical Sensors and Sensing Congress 2022 (AIS, LACSEA, Sensors, ES)*, Washington, D.C.: Optica Publishing Group, ISBN: 978-1-957171-10-4. DOI: 10.1364/AIS.2022.ATu3G.2.

[181] J. Goldschmidt, E. Moser, L. Nitzsche, R. Bierl, and J. Wöllenstein, "Approaches to overcome data scarcity when utilizing artificial neural networks in quantitative gas analysis," *tm - Technisches Messen*, 2023. DOI: doi:10.1515/teme-2023-0051. [Online]. Available: https://doi.org/10.1515/teme-2023-0051.

[182] J. J. Haworth, C. K. Pitcher, G. Ferrandino, A. R. Hobson, K. L. Pappan, and J. L. D. Lawson, "Breathing new life into clinical testing and diagnostics: Perspectives on volatile biomarkers from breath," *Critical reviews in clinical laboratory sciences*, vol. 59, no. 5, pp. 353–372, 2022. DOI: 10.1080/10408363.2022.2038075.

[183] S. Weigl, Ed., *Breath Analysis* (Bioanalytical Reviews). Cham: Springer International Publishing, 2023, ISBN: 978-3-031-18525-0. DOI: 10.1007/978-3-031-18526-7.

[184] B. Henderson, A. Khodabakhsh, M. Metsälä, *et al.*, "Laser spectroscopy for breath analysis: Towards clinical implementation," *Applied Physics B*, vol. 124, no. 8, p. 161, 2018, ISSN: 0946-2171. DOI: 10.1007/s00340-018-7030-x.

[185] A. Prabhakar, A. Quach, H. Zhang, *et al.*, "Acetone as biomarker for ketosis buildup capability–a study in healthy individuals under combined high fat and starvation diets," *Nutrition journal*, vol. 14, p. 41, 2015. DOI: 10.1186/s12937-015-0028-x.

[186] K. Musa-Veloso, S. S. Likhodii, and S. C. Cunnane, "Breath acetone is a reliable indicator of ketosis in adults consuming ketogenic meals," *The American journal of clinical nutrition*, vol. 76, no. 1, pp. 65–70, 2002, ISSN: 0002-9165. DOI: 10.1093/ajcn/76.1.65.

[187] K. Musa-Veloso, S. S. Likhodii, E. Rarama, *et al.*, "Breath acetone predicts plasma ketone bodies in children with epilepsy on a ketogenic diet," *Nutrition (Burbank, Los Angeles County, Calif.)*, vol. 22, no. 1, pp. 1–8, 2006, ISSN: 0899-9007. DOI: 10.1016/j.nut.2005.04.008.

[188] M. A. Samara, W. H. W. Tang, F. Cikach, *et al.*, "Single exhaled breath metabolomic analysis identifies unique breathprint in patients with acute decompensated heart failure," *Journal of the American College of Cardiology*, vol. 61, no. 13, pp. 1463–1464, 2013. DOI: 10.1016/j.jacc.2012.12.033.

[189] T. P. J. Blaikie, J. A. Edge, G. Hancock, *et al.*, "Comparison of breath gases, including acetone, with blood glucose and blood ketones in children and adolescents with type 1 diabetes," *Journal of breath research*, vol. 8, no. 4, 2014. DOI: 10.1088/1752-7155/8/4/046010.

[190] P. R. Galassetti, B. Novak, D. Nemet, *et al.*, "Breath ethanol and acetone as indicators of serum glucose levels: An initial report," *Diabetes technology & therapeutics*, vol. 7, no. 1, pp. 115–123, 2005, ISSN: 1520-9156. DOI: 10.1089/dia.2005.7.115.

[191] A. Reyes-Reyes, R. C. Horsten, H. P. Urbach, and N. Bhattacharya, "Study of the exhaled acetone in type 1 diabetes using quantum cascade laser spectroscopy," *Analytical chemistry*, vol. 87, no. 1, pp. 507–512, 2015. DOI: 10.1021/ac504235e.

[192] V. Ruzsányi and M. Péter Kalapos, "Breath acetone as a potential marker in clinical practice," *Journal of breath research*, vol. 11, no. 2, 2017. DOI: 10.1088/1752-7163/aa66d3.

[193] Y. Yuan, Z. Chen, X. Zhao, *et al.*, "Continuous monitoring of breath acetone, blood glucose and blood ketone in 20 type 1 diabetic outpatients over 30 days," *Journal of Analytical & Bioanalytical Techniques*, vol. 08, no. 05, 2017. DOI: 10.4172/2155-9872.1000386.

[194] K. L. Moskalenko, A. I. Nadezhdinskii, and I. A. Adamovskaya, "Human breath trace gas content study by tunable diode laser spectroscopy technique," *Infrared physics & technology*, vol. 37, no. 1, pp. 181–192, 1996. DOI: 10.1016/1350-4495(95)00097-6.

[195] K. Musa-Veloso, S. S. Likhodii, E. Rarama, *et al.*, “Breath acetone predicts plasma ketone bodies in children with epilepsy on a ketogenic diet,” *Nutrition*, vol. 22, no. 1, pp. 1–8, 2006. DOI: 10.1016/j.nut.2005.04.008.

[196] A. M. Diskin, P. Španěl, and D. Smith, “Time variation of ammonia, acetone, isoprene and ethanol in breath: A quantitative sift-ms study over 30 days,” *Physiological measurement*, vol. 24, no. 1, p. 107, 2003. DOI: 10.1088/0967-3334/24/1/308.

[197] K. Schwarz, A. Pizzini, B. Arendacka, *et al.*, “Breath acetone—aspects of normal physiology related to age and gender as determined in a ptr-ms study,” *Journal of Breath Research*, vol. 3, no. 2, 2009. DOI: https://doi.org/10.1088/1752-7155/3/2/027003.

[198] J. Beauchamp, C. Davis, and J. Pleil, *Breathborne Biomarkers and the Human Volatilome*. Elsevier, 2020, ISBN: 9780128199671. DOI: 10.1016/C2018-0-04980-4.

[199] C. Turner, P. Španěl, and D. Smith, “A longitudinal study of ethanol and acetaldehyde in the exhaled breath of healthy volunteers using selected-ion flow-tube mass spectrometry,” *Rapid Communications in Mass Spectrometry*, vol. 20, no. 1, pp. 61–68, 2006. DOI: 10.1002/rcm.2275.

[200] P. Španel, K. Dryahina, and D. Smith, “The concentration distributions of some metabolites in the exhaled breath of young adults,” *Journal of Breath Research*, vol. 1, no. 2, 2007. DOI: 10.1088/1752-7155/1/2/026001.

[201] A. Mazzatenta, C. Di Giulio, and M. Pokorski, “Pathologies currently identified by exhaled biomarkers,” *Respiratory Physiology & Neurobiology*, vol. 187, no. 1, pp. 128–134, 2013. DOI: 10.1016/j.resp.2013.02.016.

[202] L. Ciaffoni, G. Hancock, J. J. Harrison, *et al.*, “Demonstration of a mid-infrared cavity enhanced absorption spectrometer for breath acetone detection,” *Analytical chemistry*, vol. 85, no. 2, pp. 846–850, 2013. DOI: 10.1021/ac3031465.

[203] R. Centeno, J. Mandon, F. Harren, and S. Cristescu, “Influence of ethanol on breath acetone measurements using an external cavity quantum cascade laser,” *Photonics*, vol. 3, no. 2, p. 22, 2016. DOI: 10.3390/photonics3020022.

[204] Y. V. Kistenev, A. V. Borisov, D. A. Kuzmin, O. V. Penkova, N. Y. Kostyukova, and A. A. Karapuzikov, “Exhaled air analysis using wideband wave number tuning range infrared laser photoacoustic spectroscopy,” *Journal of biomedical optics*, vol. 22, no. 1, 2017. DOI: 10.1117/1.JBO.22.1.017002.

[205] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, “Emcee: The mcmc hammer,” *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 925, p. 306, 2013. DOI: 10.1086/670067.

[206] K. Shariat, D. Kirsanov, A. C. Olivieri, and H. Parastar, “Sensitivity and generalized analytical sensitivity expressions for quantitative analysis using convolutional neural networks,” *Analytica chimica acta*, vol. 1192, 2022. DOI: 10.1016/j.aca.2021.338697.

[207] T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim, "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 1, pp. 35–48, 2009, ISSN: 01697439. DOI: 10.1016/j.chemolab.2008.08.004.

[208] A. Zifarelli, A. Cantatore, A. Sampaolo, *et al.*, "Multivariate analysis and digital twin modelling: Alternative approaches to molecular relaxation in photoacoustic spectroscopy. (under review)," *Photoacoustics*,

[209] Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., Ed., *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate.* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2007.

[210] J. T. Friedlein, E. Baumann, K. A. Briggman, *et al.*, "Dual-comb photoacoustic spectroscopy," *Nature communications*, vol. 11, no. 1, p. 3152, 2020. DOI: 10.1038/s41467-020-16917-y.

## Co-authored Publications

"Effects of ambient parameters and cross-sensitivities from O2, CO2 and H2O on the photoacoustic detection of acetone in the UV region",
Stefan Weigl, Elisabeth Wittmann, Thomas Rück, Rudolf Bierl, Frank-Michael Matysik
*Sensors and Actuators B: Chemical*, vol. 328, 2021,
https://doi.org/10.1016/j.snb.2020.129001.

"Using a Modulated Quantum Cascade Laser for Photoacoustic Spectra Recording of Exhaled Acetone and Main Breath Components",
Jonas Pangerl, Elisabeth Wittmann, Stefan Weigl, Max Müller, Rudolf Bierl, and Frank-Michael Matysik
*Optical Sensors and Sensing Congress 2022*, Paper ATu3G.1, 2022,
https://doi.org/10.1364/AIS.2022.ATu3G.1.

"Luminance Simulation in CARLA Under Cloud Coverage - Model Validation and Implications",
Fabian Ulreich, Elisabeth Moser, Florian Olbrich, Martin Ebert, Rudolf Bierl, Andre Kaup
in *2023 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, pp. 228-233, 2023
https://doi.org/10.1109/MetroAutomotive57488.2023.10219098

"Semi-Selective Array for the Classification of Purines with Surface Plasmon Resonance Imaging and Deep Learning Data Analysis",
Simon Jobst, Patrick Recum, Ángela Écija-Arenas, Elisabeth Moser, Rudolf Bierl, Thomas Hirsch
*ACS Sensors*, 2023,
https://doi.org/10.1021/acssensors.3c01114.

## Poster presentations

"Argo: Towards Small Vessel Detection for Humanitarian Purposes ",
Elisabeth Moser, Selina Meyer, Maximilian Schmidhuber, Daniel Ketterer, Matthias Eberhardt
Presented at the *Thirty-First International Joint Conference on Artificial Intelligence* in 2022 (Vienna, Austria),
https://doi.org/10.24963/ijcai.2022/728.

## Oral presentation

"AI for Breath Analysis",
Elisabeth Moser
Presented at the *KI Campus Ostbayern: AI Talks* in 2022 (Regensburg, Germany)

"Konzepte und Anwendungen für die automatisierte Auswertung von Infrarot-Gasspektren mittels KI",
Elisabeth Moser und Jens Goldschmidt
Presented at the *VDI-Expertenforum: Mit Intelligenz von der Messung zur Information* in 2022 (Karlsruhe, Germany)

## The authors' original publications

Parts that were adapted from the author's publications form the basis of the result section III. This section lists the abstracts of the original publications.

**A deep learning system to transform cross-section spectra to varying environmental conditions**

Elisabeth Moser, Simon Jobst, Rudolf Bierl, and Frank Jenko

in *Vibrational Spectroscopy*, Volume 122, 2022, https://doi.org/10.1016/j.vibspec.2022.103410.

**Abstract** Absorption cross-sections provide a basis for many gas sensing applications. Therefore, any error in molecular cross-sections caused by varying environmental conditions propagates to spectroscopic applications. Original molecular cross-sections in varying environmental conditions can only be simulated for some molecules, whereas for most multi-atom molecules, one must rely on high-precision measurements at certain environmental configurations. In this study, a deep learning system trained with simulated absorption cross-sections for predicting cross-sections at a different pressure configuration is presented. The system's capability to transfer to measured, multi-atom cross-sections is demonstrated. Thus, it provides an alternative to (pseudo-) line lists whenever the required information for simulation is unavailable. The predictive performance of the system was evaluated on validation data via simulation, and its transfer learning capabilities were demonstrated on actual measurement chlorine nitrate data. From the comparison between the system and line lists, the system shows slightly worse performance than pseudo-line lists but its predictive quality is still deemed acceptable with less than 5 % relative integral change with a highly localized error around the peak center. This opens a promising way for further research to use deep learning to simulate the effect of varying environmental conditions on absorption cross-sections.

**Modeling the Photoacoustic Spectrum of a Quantum Cascade Laser for Human Breath**

Elisabeth Moser, Jonas Pangerl, Simon Jobst, Stefan Weigl, and Rudolf Bierl

at *Optical Sensors and Sensing Congress 2022*, Paper ATu3G.2,
https://doi.org/10.1364/AIS.2022.ATu3G.2.

**Abstract** A modeling approach to create a photoacoustic spectrum from synthetic data is presented and evaluated. The resulting model reaches a MAPE score of 2.7% and can be used to enable data-driven development in future work.

**A sub-ppbv-level Acetone and Ethanol Quantum Cascade Laser Based Photoacoustic Sensor – Characterization and Multi-Component Spectra Recording in Synthetic Breath**

Jonas Pangerl and Elisabeth Moser (shared first authorship), Max Müller, Stefan Weigl, Simon Jobst, Thomas Rück, Rudolf Bierl, Frank-Michael Matysik

in *Photoacoustics*, vol. 30, 2022, https://doi.org/10.1016/j.pacs.2023.100473.

**Abstract**

Trace gas analysis in breath is challenging due to the vast number of different components. We present a highly sensitive quantum cascade laser based photoacoustic setup for breath analysis. Scanning the range between 8263 and 8270 nm with a spectral resolution of 48 pm, we are able to quantify acetone and ethanol within a typical breath matrix containing water and CO2. We photoacoustically acquired spectra within this region of mid-infra-red light and prove that those spectra do not suffer from non-spectral interferences. The purely additive behavior of a breath sample spectrum was verified by comparing it with the independently acquired single component spectra using Pearson and Spearman correlation coefficients. A previously presented simulation approach is improved and an error attribution study is presented. With a 3$\sigma$ detection limit of 6.5 ppbv in terms of ethanol and 250 pptv regarding acetone, our system is among the best performing presented so far.

**Improving the Performance of Artificial Neural Networks Trained on Synthetic Data in Gas Spectroscopy – a Study on Two Sensing Approaches**

Jens Goldschmidt and Elisabeth Moser (shared first authorship), Leonard Nitzsche, Rudolf Bierl, Jürgen Wöllenstein

in *Technisches Messen*, vol. 90, 2022, https://doi.org/10.1515/teme-2023-0051

**Abstract**

Artificial neural networks (ANNs) are used in quantitative infrared gas spectroscopy to predict concentrations on multi-component absorption spectra. Training of ANNs requires vast amounts of labelled training data which may be elaborate and time consuming to obtain. Additional data can be gained by the utilization of synthetically generated spectra, but at the cost of systematic deviations to measured data. Here, we present two approaches to train ANNs with a combination of comparatively small, measured data sets and synthetically generated data. For the first approach a neural network is trained hybridly with synthetically generated infrared absorption spectra of mixtures of N2O and CO and measured zero-gas spectra, taken with a mid-infrared dual comb spectrometer. This improves the mean absolute error (MAE) of the network predictions from 0.46 to 0.01 ppmV and 0.24 to 0.01 ppmV for the concentration predictions of N2O and CO respectively for zero-gas measurements which was previously observed for training with purely synthetic data. At the same time a similar performance on spectra from gas mixtures of 0 to 100 ppmV N2O and 0 to 60 ppmV CO was achieved. For the second approach an ANN pre-trained on synthetic infrared spectra of mixtures of acetone and ethanol is retrained on a small dataset consisting of 26 spectra taken with a mid-infrared photoacoustic spectrometer. In this case the MAE for the concentration predictions of ethanol and acetone are improved by 45% and 20% in comparison to purely synthetic training. This shows the capability of using synthetically generated data to train ANNs in combination with small amounts of measured data to further improve neural networks for gas sensing and the transferability between different sensing approaches.

## Declaration of Collaboration

Most of the theoretical and experimental scientific work that is presented within this thesis was done independently by the author. In some cases, however, the practical implementation of concepts and the performance of measurements was carried out in collaboration with other researchers and individuals. In any case, assistance was guided and supervised by the author.

***Quantum cascade laser (QCL) setup and wavelength characterisation (Chapter 7.1)***
The integration into the measurement station and the optical alignment of the QCL have been performed by Florian Feldmeier and Jonas Pangerl. The measurement routine was developed by Stefan Weigl in close consultation with the author and Jonas Pangerl. The data collection was carried out by the author and Jonas Pangerl. The system specification was performed by Jonas Pangerl in close consultation with the author and Stefan Weigl.

***Interband cascade laser (ICL) setup and data acquisition (Chapter 7.2)***
Andrea Zifarelli and Angelo Sampaolo provided the QEPAS measurement data. Aldo Cantatore performed the measurements.

***Kernel Size Experiments for Spectral Convolutional Networks (Section 6.2.2)***
Experiments to determine the effect of kernel size on model performance were conducted by Nikhitha Gudur under the supervision and in close consultation with the author.

***Disentangled Autoencoder Experiments (Section 6.2.1)***
Tristan Schaar conducted Disentangled Autoencoder experiments under the supervision and in close consultation with the author. The author performed the interpretation of the results in context with this thesis.

***Wavelength Modulation Simulation (Section 7.2.1)***
The simulation of the wavelength modulation was created in close consultation with Max Müller.

***CoNRad (Section 7.2.1)***
The CoNRad algorithm was programmed and verified by Max Müller in collaboration with Thomas Rück and Simon Jobst.

***Grammar-Checking***

Grammarly was used for spell-checking and to improve grammar and readability.

**Funding**

Financial support was received within the scope of the *BreathSens* research project by the Federal Ministry of Education and Research (BMBF) - Grant number: 13GW0325C.

Financial support was received within the scope of the *AdlabSens* research project by the Bavarian Ministry of Economic Affairs, Regional Development and Energy - Grant number: DIK0230/02.

The author received financial support from the BayWISS - Digitization initiative from September 2020 to September 2023.

The author received financial support from the Hanns-Seidel-Foundation from September 2020 to September 2023.

The author received financial support from the MLP Finanzberatung - SE in October 2023.

# List of Acronyms

**2f** Second Harmonic

**AAE** Adversarial Autoencoder

**ACS** Absorption Cross-Section

**AE** Autoencoder

**AI** Artificial Intelligence

**am** Amplitude Modulation

**am-PAS** Amplitude Modulated Photoacoustic Spectroscopy

**ANN** Artificial Neural Network

**API** Application Programming Interface

**AREP** Average Relative Error of Prediction

**CNN** Convolutional Neural Network

**CoNRad** Algorithm to Compute the Collision Based Non-radiative Efficiency and Phase Lag of Energy Relaxation on a Molecular Level

**COVID-19** Coronavirus Disease 2019

**CovSel** Covariance Selection

**cw** Continous Wave

**DFB** Distributed Feedback Laser

**DOF** Degree of Freedom

**DT** Decision Tree

**EPLL** Estimated-Pseudo Line List

**GAN** Generative Adversarial Network

**GB** Gradient Boosting

**ICL** Interband Cascade Laser

**IR** Infrared

**JPL** Jet Propulsion Laboratory

| | |
|---|---|
| **FFT** | Fast-Fourier Transformation |
| **fIR** | Far-infrared |
| **FPGA** | Field Programmable Gate Array |
| **FTIR** | Fourier Transform Infrared |
| **FWHM** | Full Width at Half Maximum |
| **hapi** | HITRAN Application Programming Interface |
| **HITRAN** | High-Resolution Transmission Molecular Absorption Database |
| **HT** | Hartmann–Tran |
| **HWHM** | Half Width at Half Maximum |
| **LD** | Laser Driver |
| **LDC** | Laser Driver Controller |
| **LIA** | Lock-in Amplifier |
| **NDIR** | Non–dispersive Infrared |
| **nIR** | Near-infrared |
| **NMR** | Nuclear Magnetic Resonance |
| **MACCS** | Molecular Access System |
| **MAE** | Mean Absolute Error |
| **MAPE** | Mean Average Percentage Error |
| **MCMC** | Markov Chain Monte Carlo |
| **MCR** | Multi Curve Regression |
| **MEMS** | Micro Electro Mechanical System |
| **mIR** | Mid-infrared |
| **ML** | Machine Learning |
| **MLR** | Multi Linear Regression |
| **MSE** | Mean Squared Error |
| **OPO** | Optical Parametric Oscillator |

| | |
|---|---|
| **PAS** | Photoacoustic Spectroscopy |
| **PCA** | Principal Component Analysis |
| **PCB** | Printed Circuit Board |
| **pCqSDHC** | Partially–Correlated Quadratic–Speed–Dependent Hard–Collision |
| **PLS** | Partial Least Squares |
| **PLSR** | Partial Least Squares Regression |
| **PNNL** | Pacific Northwest National Laboratory |
| **QCL** | Quantum Cascade Laser |
| **QEPAS** | Quartz-Enhanced Photoacoustic Spectroscopy |
| **QTF** | Quartz Tuning Fork |
| **ReLu** | Rectified Linear Unit |
| **RF** | Random Forest |
| **RIC** | Relative Integral Change |
| **RMSE** | Root Mean Squared Error |
| **SE** | Squeeze and Excite |
| **SGD** | Stochastic Gradient Descend |
| **SVM** | Support Vector Machine |
| **SVR** | Support Vector Regression |
| **TDLAS** | Tunable Diode Laser Absorption Spectroscopy |
| **TEC** | Temperature Controller |
| **UV** | Ultraviolet |
| **VAE** | Variational Autoencoder |
| **VIS** | Visible |
| **wm** | Wavelength Modulation |
| **wm-PAS** | Wavelength Modulated Photoacoustic Spectroscopy |

# List of Symbols

| Constants | | Unit |
|---|---|---|
| $c_0$ | Speed of Light in Vacuum = 299 792 458 | $m\,s^{-1}$ |
| h | Planck's Quantum of Action = $6.626 \cdot 10^{-34}$ | J s |
| k | Boltzmann Constant = $1.380649 \cdot 10^{-23}$ | $J\,K^{-1}$ |
| $N_A$ | Avogadro Constant = $6.02214076 \cdot 10^{23}$ | $mol^{-1}$ |
| R | Ideal Gas Constant = 8.134 472 | J/mol K |
| $T_{ref}$ | Reference Temperature = 296K | K |

| Greek Symbols | | Unit |
|---|---|---|
| $\alpha(\tilde{\nu})$ | Absorption Coefficient at $\tilde{\nu}$ | $cm^{-1}$ |
| $\Delta\tilde{\nu}_{col}$ | FWHM for Collision Broadening | $cm^{-1}$ |
| $\Delta\tilde{\nu}_{dicke}$ | FWHM for Dicke Narrowing | $cm^{-1}$ |
| $\Delta\tilde{\nu}_{dop}$ | FWHM for Doppler Broadening | $cm^{-1}$ |
| $\Delta\tilde{\nu}_{nat}$ | FWHM for Natural Broadening | $cm^{-1}$ |
| $\Delta$ | Laplace Operator: Second Partial Derivative in Space | $m^{-2}$ |
| $\Delta_0$ | Speed–Averaged Line–Shift | $cm^{-1}$ |
| $\Delta_2$ | Speed Dependence of the Line–Shift | $cm^{-1}$ |
| $\Delta_{shift}$ | Frequency Shift | $s^{-1}$ |
| $\epsilon(\tilde{\nu})$ | Molar Extinction Coefficient at $\tilde{\nu}$ | $L\,mol^{-1}\,cm^{-1}$ |
| $\epsilon_{relax}$ | Relaxation Efficiency | 1 |
| $\eta$ | Line Shape Correlation Parameter | 1 |
| $\gamma$ | Heat Capacity Ratio | 1 |
| $\gamma_0$ | HWHM for Collision Broadening | $cm^{-1}$ |
| $\gamma_2$ | Speed Dependence of the Line–Width | $cm^{-1}$ |
| $\gamma_D$ | HWHM for Doppler Broadening | $cm^{-1}$ |
| $\lambda_{ph}$ | Wavelength of a Photon | nm |
| $\epsilon_0$ | Vacuum Permittivity Constant | A s m |
| $\mu$ | Transition Dipole Moment | A s m |
| $\mu_j$ | Reciprocal Light to Sound Coupling Factor of the j-th Mode | 1 |
| $\mu_{ij}$ | Reduced Mass of Two Molecules i and j | kg |
| $\nu_{ph}$ | Frequency of a Photon | $s^{-1}$ |
| $\omega$ | Angular Frequency | rad/s |
| $\phi$ | Phase Lag | 1 |
| $\rho$ | Particle Density | $cm^{-3}$ |
| $\rho$ | Volume Number Density | $molecules/m^3$ |
| $\rho_i$ | Particle Density of the Molecular State i | $cm^{-3}$ |
| $\sigma_A(\tilde{\nu})$ | Absorption Cross–Section at $\tilde{\nu}$ | $cm^2/molecule$ |

| | | |
|---|---|---|
| $\sigma_{col}$ | Collision Cross-Section | $\mathrm{m}^{-2}$ |
| $\tau_f$ | Total Lifetime Constant of an Excited State $v$ | s |
| $\tau_i$ | Lifetime of a Molecular State i | s |
| $\tau_n$ | Non-Radiative Lifetime Constant of an Excited State | s |
| $\tau_r$ | Radiative Lifetime Constant of an Excited State | s |
| $\tau_{col}$ | Collisional Lifetime | s |
| $\tilde{\nu}$ | Wavenumber | $\mathrm{cm}^{-1}$ |
| $\tilde{\nu}_0$ | Wavenumber of the Line Center | $\mathrm{cm}^{-1}$ |
| $\tilde{\nu}_{\mathrm{ph}}$ | Wavenumber of a Photon | $\mathrm{cm}^{-1}$ |

| **Latin Symbols** | | **Unit** |
|---|---|---|
| $w\,(x, y)$ | the Complex Probability Function | 1 |
| $\dot{H}(t)$ | Power Density | $\mathrm{W\,m}^{-2}$ |
| $\dot{H}_0$ | Amplitude Power Density at Equilibrium | $\mathrm{W\,m}^{-2}$ |
| $\mathcal{N}$ | Gaussian Probability Density Function | 1 |
| $\overline{v}_{rel}$ | Relative Mean Velocity | $\mathrm{m\,s}^{-1}$ |
| $\tilde{A}(\tilde{\nu})$ | Absorbance at $\tilde{\nu}$ | 1 |
| $\tilde{B}$ | Rotational Constant | $\mathrm{cm}^{-1}$ |
| $\tilde{D}_e$ | Depth of the Potential Well | $\mathrm{cm}^{-1}$ |
| $\tilde{g}_i$ | Statistical Weight of Molecular State i | 1 |
| $\tilde{T}(\tilde{\nu})$ | Transmittance at $\tilde{\nu}$ | 1 |
| $\tilde{x}$ | Separation from Shifted Transition Frequency | 1 |
| $\tilde{y}_i$ | Target Variable of an Algorithm | 1 |
| $\vec{r}$ | Three Dimensional Position in Space | 1 |
| $A$ | Einstein Coefficient of Spontaneous Emission | $\mathrm{s}^{-1}$ |
| $A'(\tilde{\nu})$ | Absorption at $\tilde{\nu}$ | 1 |
| $a_M$ | Displacement Factor | $\mathrm{m}^{-1}$ |
| $A_{\tilde{\nu}c}$ | Absorption Spectra over the Wavelength Range Scaled by Concentration | 1 |
| $B'$ | Einstein Coefficient of Stimulated Emission | $\mathrm{m}^3\,\mathrm{J}^{-1}\,\mathrm{s}^{-2}$ |
| $B_{nm}$ | Einstein Coefficient of Absorption from State n to m | $\mathrm{m}^3\,\mathrm{J}^{-1}\,\mathrm{s}^{-2}$ |
| $c$ | Concentration | $ppmV$ |
| $C\,(t)$ | Correlation Function of Radiating Dipole | 1 |
| $c_i$ | Concentration of Molecule i | $\mathrm{mol\,L}^{-1}$ |
| $C_p$ | Isobar Heat Capacity | $\mathrm{J\,K}^{-1}$ |
| $c_s$ | Speed of Sound | $\mathrm{m\,s}^{-1}$ |
| $C_V$ | Isochore Heat Capacity | $\mathrm{J\,K}^{-1}$ |
| $C_{cell}$ | Cell Constant | |
| $E_e$ | Energy of an Electonic Transition | J |
| $E_i$ | Energy of a Molecular State i | J |
| $E_r$ | Energy of an Rotational Transition | J |
| $E_v$ | Energy of an Vibrational Transition | J |

| | | |
|---|---|---|
| $E_{\mathrm{ph}}$ | Energy of a Photon | J |
| $f_{LS}$ | Line Shape Function | cm |
| $G$ | Vibrational Term of Line Shape Function | $\mathrm{cm^{-1}}$ |
| $g_col\left(\tilde{\nu}\right)$ | Unnormalized Line Shape Function for Collision Broadening | 1 |
| $g_{nat}\left(\tilde{\nu}\right)$ | Unnormalized Line Shape Function for Natural Broadening | 1 |
| $I$ | Moment of Inertia | $\mathrm{kg\,m^{2}}$ |
| $I_0(\tilde{\nu})$ | Incident Light Intensity at Wavenumber $\tilde{\nu}$ | $\mathrm{W\,m^{-2}}$ |
| $I_1(\tilde{\nu})$ | Transmitted Light Intensity at Wavenumber $\tilde{\nu}$ | $\mathrm{W\,m^{-2}}$ |
| $I_D$ | Lineshape Function for the Doppler Profile | 1 |
| $I_h$ | High Level Current Supplied to the Laser during Square Modulation | A |
| $I_L$ | Lineshape Function for the Lorentz Profile | 1 |
| $I_V$ | Lineshape Function for the Voigt Profile | 1 |
| $I_{pCqSDHC}$ | Lineshape Function for the Hartmann-Tran Profile | 1 |
| $j$ | Angular Momentum Quantum Number | 1 |
| $K\left(x,y\right)$ | the Voigt Function | 1 |
| $k_v$ | Spring Force | $\mathrm{N\,m^{-1}}$ |
| $L_R$ | Length of the Acoustic Resonator | m |
| $l_{opt}$ | Length of the Optical Path | m |
| $m$ | Mass | kg |
| $M_i$ | Molar Mass of Molecule i | kg/mol |
| $m_{eff}$ | Effective Mass | kg |
| $m_i$ | Mass of Molecule i | kg |
| $n$ | Number of Samples | 1 |
| $N_i$ | Volume Ratio of i | $\mathrm{molecules/cm^{3}}$ |
| $N_{atoms}$ | Number of Atoms in a Molecule | 1 |
| $p_a$ | Pressure Amplitude Induced by Light Absorption | $Pa$ |
| $p_j$ | Normalization Coefficient of Mode j | 1 |
| $p_j(\vec{r})$ | Spatial Pressure Distribution of Mode j | 1 |
| $P_o$ | Optical Power | W |
| $p_{\mathrm{pas}}$ | Synthetic Photoacoustic Signal Estimation after Amplification | µV |
| $P_{\tilde{\nu}}$ | Relative Laser Output Power over the Wavelength Range | 1 |
| $p_{MW}\left(v\right)$ | Probability of a Velocity by the Maxwell–Boltzmann Distribution | 1 |
| $Q_j$ | Quality Factor of the Mode | 1 |
| $Q_{int}(T)$ | Total Internal Partition Function of a Molecule | 1 |
| $r_i$ | Radius of Molecule i | m |
| $s$ | Line Strength | cm/molecule |
| $T$ | Temperature | K |
| $t$ | Time | s |
| $V$ | Potential Energy | J |
| $v$ | Velocity | $\mathrm{m\,s^{-1}}$ |
| $v$ | Vibrational Quantum Number | 1 |
| $V_m$ | Molar Volume | $\mathrm{L\,mol^{-1}}$ |

| | | |
|---|---|---|
| $V_R$ | Volume of the Acoustic Resonator | $\mathrm{m}^3$ |
| $\nu_{VC}$ | Velocity–Changing Frequency | $\mathrm{cm}^{-1}$ |
| $x$ | Separation from Unperturbed Transition Frequency | 1 |
| $x_e$ | Anharmonic Constant | 1 |
| $x_i$ | Input Variable of an Algorithm | 1 |
| $x_r$ | Displacement of the Molecular Bond Length from Equilibrium | m |
| $y$ | Effective Broadening Parameter | 1 |
| $y_i$ | Prediction Variable of an Algorithm | 1 |

# List of Figures

# List of Tables