atmosphere

MDPI

Article

High-Performance VOC Quantification for IAQ Monitoring Using Advanced Sensor Systems and Deep Learning

Yannick Robin *, Johannes Amann, Tobias Baur, Payman Goodarzi, Caroline Schultealbert, Tizian Schneider and Andreas Schütze

Lab for Measurement Technology, Saarland University, Campus A5 1, 66123 Saarbrücken, Germany; j.amann@lmt.uni-saarland.de (J.A.); t.baur@lmt.uni-saarland.de (T.B.); p.goodarzi@lmt.uni-saarland.de (P.G.); c.schultealbert@lmt.uni-saarland.de (C.S.); t.schneider@lmt.uni-saarland.de (T.S.); schuetze@lmt.uni-saarland.de (A.S.)

* Correspondence: y.robin@lmt.uni-saarland.de

Abstract: With air quality being one target in the sustainable development goals set by the United Nations, accurate monitoring also of indoor air quality is more important than ever. Chemiresistive gas sensors are an inexpensive and promising solution for the monitoring of volatile organic compounds, which are of high concern indoors. To fully exploit the potential of these sensors, advanced operating modes, calibration, and data evaluation methods are required. This contribution outlines a systematic approach based on dynamic operation (temperature-cycled operation), randomized calibration (Latin hypercube sampling), and the use of advances in deep neural networks originally developed for natural language processing and computer vision, applying this approach to volatile organic compound measurements for indoor air quality monitoring. This paper discusses the pros and cons of deep neural networks for volatile organic compound monitoring in a laboratory environment by comparing the quantification accuracy of state-of-the-art data evaluation methods with a 10-layer deep convolutional neural network (TCOCNN). The overall performance of both methods was compared for complex gas mixtures with several volatile organic compounds, as well as interfering gases and changing ambient humidity in a comprehensive lab evaluation. Furthermore, both were tested under realistic conditions in the field with additional release tests of volatile organic compounds. The results obtained during field testing were compared with analytical measurements, namely the gold standard gas chromatography mass spectrometry analysis based on Tenax sampling, as well as two mobile systems, a gas chromatograph with photo-ionization detection for volatile organic compound monitoring and a gas chromatograph with a reducing compound photometer for the monitoring of hydrogen. The results showed that the TCOCNN outperforms state-of-the-art data evaluation methods, for example for critical pollutants such as formaldehyde, achieving an uncertainty of around 11 ppb even in complex mixtures, and offers a more robust volatile organic compound quantification in a laboratory environment, as well as in real ambient air for most targets.

Keywords: volatile organic compounds (VOCs); indoor air quality (IAQ); deep neural networks; neural network architecture search; temperature-cycled operation (TCO)

Citation: Robin, Y.; Amann, J.; Baur, T.; Goodarzi, P.; Schultealbert, C.; Schneider, T.; Schütze, A. High-Performance VOC Quantification for IAQ Monitoring Using Advanced Sensor Systems and Deep Learning. *Atmosphere* **2021**, *12*, 1487. https://doi.org/10.3390/atmos12111487

Academic Editor: Stéphane Le Calvé

Received: 15 October 2021
Accepted: 5 November 2021
Published: 10 November 2021

1. Introduction

With indoor air quality (IAQ) being one of the most common and unavoidable threats to human health and also one of the most difficult to determine accurately, it is more important than ever to be able to make accurate measurements of IAQ [1]. Especially dangerous are volatile organic compounds (VOCs), which can lead to serious health problems. For example, extensive exposure to formaldehyde can cause cancer [2]. Even the United Nations agree in their goals for sustainable development that pollution is a goal of the greatest importance and that the number of deaths and illnesses from hazardous chemicals and air, water, and soil pollution and contamination should be substantially

reduced by 2030 [3]. Several reasons are making an accurate measurement of IAQ difficult. First of all, indoor air contains hundreds or even thousands of compounds, some of them benign, others toxic, even at very low concentrations, making accurate quantification of each of them impossible, at least for routine and continuous measurements [4]. Second, analytical measurement systems that are capable of providing measurements of the most relevant VOCs are very expensive, are too slow for real-time application, and require expert knowledge to operate and calibrate [5]. Third, due to the difficulty of providing comprehensive measurements, too little is known about the cause and effect of various gases and especially of their combined effect [6]. Currently, CO_2 is the primary indicator used for IAQ estimation as there is a direct relation between VOC concentration in room air and the CO_2 concentration if the VOC levels are caused by human presence, as already described by Pettenkofer in 1858 [7]. However, dangerous VOCs are also released from building materials, furniture, and activities such as cooking and cleaning, which do not release CO_2 [8–10]. For this study, VOCs represent a diverse spectrum from very volatile (VVOC) to semivolatile (SVOC) organic compounds [11]. In this study, we concentrated on VOCs with a high-to-medium vapor pressure including the carcinogens formaldehyde and benzene, which are considered as two of the most toxic substances in indoor air with guideline threshold values in the low ppb range according to the WHO [11]. Therefore, comprehensive VOC monitoring is required to provide a universal indicator for IAQ, e.g., as a basis for demand-controlled ventilation to reduce the overall burden on people [12].

We recently reported a new approach for IAQ monitoring based on low-cost metal oxide semiconductor (MOS) gas sensors (chemiresistor) combined with temperature-cycled operation (TCO) and pattern recognition to interpret the resulting complex response patterns [11,13]. In these studies, we used linear machine-learning (ML) models based on feature extraction followed by feature selection and finally regression (FESR model) to predict the concentration of various VOCs and other relevant gases individually, as well as the sum concentration of all VOCs [13]. As deep learning has proven to be very successful for the interpretation of complex patterns [14], this study provides a first test of deep-learning-based methods utilizing advanced ML techniques such as convolutional neural networks (CNNs) [15] in combination with neural architecture search (NAS) [16] for improved IAQ monitoring.

Previous studies have also successfully addressed the combination of gas sensors and deep learning [17–23]. Most of these studies have addressed higher concentrations in the ppm range [19–22] and were based on multisensor arrays [17,18,21]. Only some also used dynamic operation, but with a simple operating mode for the gas sensor with two temperatures only [19,20,22]. In some studies, the evaluation target was limited to the classification of different gases [19,23]. For a more complete overview, the reader is referred to a recent review paper on smart gas sensing technologies [24].

Therefore, the goal of this study is to show that this new deep-learning model for gas sensors should be capable of making accurate and reliable predictions for the concentration of multiple VOCs in indoor air, again based on the raw data obtained from a low-cost MOS sensor system using TCO to improve their selectivity, sensitivity, and stability [25]. Furthermore, we wanted to confirm that these models can outperform the predictions of the benchmark [13] (established linear data-driven models) at the ppb level in the laboratory environment and field tests. The benchmark was based on classic statistical approaches such as linear segmentation, principal component analysis, and a partial least-squares regression (PLSR). Finally, we compared the predictions of the deep-learning model with state-of-the-art analytical measurement systems, which are the gold standard for IAQ monitoring. Ideally, the novel approach should be considerably less costly, but able to provide high-quality data with high temporal resolution while requiring less expert knowledge, thus being easier to use.

The dataset used throughout this study was published by Baur et al. [13], and the results of the corresponding publication were used as a reference. The dataset was based on an SGP30 sensor (Sensirion AG, Stäfa, Switzerland) with four gas-sensitive layers [26],

operated using TCO for improved selectivity, sensitivity, and stability. The sensor was lab-calibrated using complex random gas mixtures [27] and then tested during operation in a typical office environment with as little human presence as possible over several weeks. Several release tests of VOCs and hydrogen were performed to validate the sensor response and to compare the performance of the model predictions of the MOS sensor system to analytical instruments [13].

2. Materials and Methods

2.1. Dataset

In order to evaluate the capabilities of the newly developed deep-learning approaches for accurate quantification of different gas concentrations in indoor air, the dataset published in Baur et al. 2021 [13] was used. This dataset utilizes advanced calibration and operation techniques together with the low-cost sensor system of the SGP30 to generate a comprehensive dataset for monitoring complex mixtures that are typical of indoor air situations. In addition to various VOCs (acetone, ethanol, formaldehyde, toluene, with formaldehyde being highly toxic, while acetone, ethanol, and toluene represent VOCs with comparatively low hazard potential), relevant inorganic gases, i.e., hydrogen and carbon monoxide, as well as relative humidity (RH), were also included in the calibration scheme, as these have a strong influence on MOS sensors (see Figure 1b). Thus, the sensors needed to be calibrated, and a machine learning model needed to be developed to discriminate interfering gases and various VOCs and to provide quantitative data on the various gas concentrations, as well as the total VOC concentration to allow comprehensive IAQ monitoring. Note that we used VOC_{sum} to describe the total VOC concentration to distinguish this from the TVOC value obtained by analytical measurements, where only VOCs with medium volatility are considered. Gas sensors, on the other hand, also detect VOCs with high volatility, so-called very volatile organic compounds (VVOCs), such as acetone, ethanol, and formaldehyde, which are not considered in the analytical TVOC value [11]. The dataset was based on random gas mixtures [27] generated in an automatic gas-mixing system [28]. With the help of this dataset, complex data-driven models for different gases can be built and evaluated in laboratory environments, as well as in real indoor air scenarios.

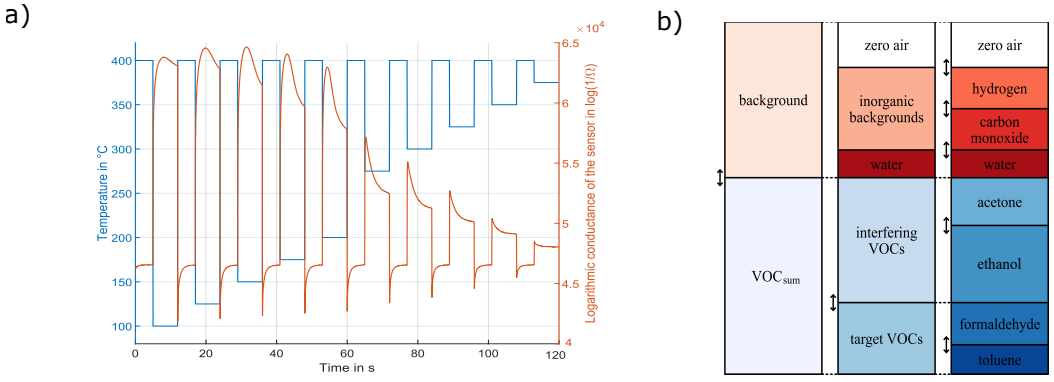


Figure 1. (**a**) Temperature-cycled operation ranging from 100–375 °C, together with one example of the logarithmic conductance of one sensor element (adapted from [13]) and (**b**) the gas composition for calibration containing background gases, as well as the target volatile organic compounds (VOCs) (adapted from [27]).

Regarding the sensor setup, the dataset utilizes an SGP30 MOS gas sensor and TCO, as illustrated in Figure 1. The sensor's output represents the resistance of the four different gas-sensitive layers over time sampled at 20 Hz. Thus, a single cycle consists of 2400 raw data samples for each of the four gas-sensing layers' resistance during TCO. This complex operation mode achieves a wide detection spectrum of the sensor system in terms of the concentration range and the gases that can be detected [29]. For improved data evaluation,

the sensor resistance patterns were converted to the logarithmic conductance of the sensor according to the Sauerwald–Baur model [30,31].

The dataset itself consists of multiple recordings of the SGP30 sensor. Those recordings can be divided into calibration phases performed in the lab with typically hundreds of well-known unique gas mixtures interlaced with field tests during which the actual gas composition and the concentrations are not known. With the help of the calibration phases, it is possible to build data-driven models for individual gases from a single sensor element. During calibration, the sensor was exposed to various gas compositions that always contained the six different gases plus relative humidity (RH), as illustrated in Figure 1, to reflect a simplified indoor environment. The concentration ranges of the various gases are given in Table 1. Furthermore, extended ranges for the VOCs as stated in Table 1 were used, to train the model also for gas compositions outside of the normally expected range, which might occur during specific exposure situations in real life, and were simulated by release tests performed in the field study (Table 2).

Table 1. Concentration ranges for all gases within gas mixtures during the calibration phases [13].

Substance	Min.	Max.	Extended
Carbon monoxide	150 ppb	2000 ppb	-
Hydrogen	400 ppb	2000 ppb	4000 ppb
Humidity	25% RH	70% RH	-
Acetone	14 ppb	300 ppb	1000 ppb
Toluene	4 ppb	300 ppb	1000 ppb
Formaldehyde	1 ppb	400 ppb	-
Ethanol	4 ppb	300 ppb	1000 ppb
VOC_{sum}	300 ppb	1200 ppb	-

Table 2. A subset of all release tests performed in [13]. Specifically listed are the release tests, which were further analyzed within this study.

Release	Time	Substance (Type of Release)	Released Amount of Substance (Approx. Increase in Room Conc.)
5	16 October, 14:50	Acetone (evaporation) Toluene (evaporation)	∼600 ppb ∼600 ppb
6	16 October, 18:00	Acetone (evaporation) Toluene (evaporation)	∼600 ppb ∼600 ppb
7	2 November, 16:50	Toluene (evaporation)	∼600 ppb
9	4 November, 16:22	Acetone (evaporation)	∼600 ppb
13	10 November, 14:30	Isopropyl alcohol (evaporation)	∼600 ppb
14	11 November, 15:49	m/p-Xylene (evaporation)	∼600 ppb
15	12 November, 15:08	Toluene (evaporation) m/p-Xylene (evaporation)	∼600 ppb ∼600 ppb
16	13 November, 14:30	Acetone (evaporation) Toluene (evaporation) Ethanol (evaporation)	∼600 ppb ∼600 ppb ∼664 ppb
17	16 November, 17:06	Hydrogen (MFC, gas cylinder)	2000 ppb

During the study, three calibration phases and two field test phases were completed (see Figure 2). The initial calibration phase and the first recalibration consisted of 100 unique gas mixtures (UGM) with the typical gas concentration ranges plus 100 additional UGM for

each of the extended concentration ranges for acetone, ethanol, toluene, and hydrogen. This resulted in a total of 500 unique gas mixtures for each of these two calibration phases. The two field test periods were performed between the calibration phases. The recalibrations were necessary to test the stability of the models, i.e., that these were still capable of reliable predictions after several weeks and that they could also suppress or compensate the drift caused by the limited stability of the gas-sensing layers. During calibration, each unique gas mixture was offered in the custom-built gas mixing apparatus for 20 min, i.e., for ten temperature cycles, as described above. Because of the nonideal synchronization between the gas-mixing system and the electronics running the temperature cycle and the delay in the gas exchange within the system, only 5 out of each 10 temperature cycles were later used for evaluation, where the gas concentration was constant. These cycles are called core samples and ensured that all measurements used for model building were recorded under stable gas compositions.

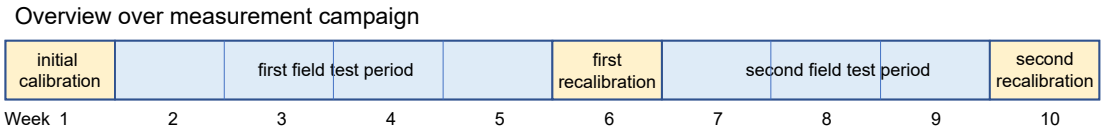


Figure 2. An illustration of the complete experiment over ten weeks, including calibration phases and field tests.

After the initial calibration, the first field test was performed in a partly controlled environment over a period of four weeks. Partly controlled means that the room was ventilated regularly and that there was no human presence in the room unless required for the release tests described below. During the field tests, the indoor air concentrations for the six trained gases were evaluated and compared with the expected values, as well as with analytical reference instrumentation for VOCs and hydrogen. To validate the quantitative prediction of the model for various gases, release tests were performed for acetone, ethanol, toluene, and hydrogen, as listed in Table 2, after thorough ventilation. Note that we did not release formaldehyde due to its high toxicity during the field tests. These tests were performed to allow an evaluation of whether the ML models can correctly detect the released compound and accurately monitor the concentration during release. The released amount was always chosen so that the concentration in the room should reach 600 ppb when the released substance was evenly distributed in the room while neglecting any losses through ventilation or adsorption on surfaces. During some of these tests, analytical instruments were used to monitor the release in parallel with the MOS sensor system. For online monitoring, a portable gas chromatograph with photo-ionization detection (GC-PID: X-pid 9500, Dräger Safety AG & Co KGaA, Lübeck, Germany) was used for VOCs and a gas chromatograph with a reducing compound photometer (GC-RCP: Peak Performer 1, Peak Laboratories LLC, Mountain View, CA, USA) for hydrogen. In addition, samples were collected on Tenax tubes (Markes International Ltd, Llantrisant, Wales, UK) for VOC monitoring in indoor air and later analyzed using thermo-desorption gas chromatography mass spectrometry (TD-GC-MS, Thermo Fisher Scientific Inc., Waltham, MA, USA). Further experimental details were given in Baur et al. [13].

2.2. Model Building

Two machine-learning approaches were used for model building. The first method, which we used as a benchmark here based on [13], utilizes feature extraction (FE) in the form of linear segmentation, standardization, feature selection (FS) based on recursive feature elimination (RFE), together with least-squares regression, a gas-mixture-based cross-validation, and an optimization scheme to find the optimum model regarding the number of selected features and the components used for the partial least-squares regression (PLSR) [13]. This approach is called Feature Extraction Selection Regression (FESR). Linear segmentation means in this case that the four different logarithmic conductance patterns obtained from the gas-sensitive layers are divided into 120 equidistant segments each and the mean and slope are calculated for all segments, resulting in a total of 960 features per

temperature cycle. The data from all unique gas mixtures offered during calibration were then split into 80% for training and 20% for testing. After this step, the 300 most important features according to the recursive feature elimination (RFE) least-squares regression (LSR) ranking were selected for further use. To find a suitable number of features and PLSR components, a gas mixture-based 10-fold cross-validation was performed on the 80% training data. Here, all core samples from 10% of the unique gas mixtures were excluded from the training for validation to find a suitable compromise for the hyperparameters, to achieve a low root-mean-squared error (RSME) with a low number of features and PLSR components. This ML model was developed using the open-source MATLAB toolbox DAV³E [32], and the approach was described in more detail in [13].

The second model-building approach then utilizes the TCOCNN architecture (see Figure 3), a 10-layer deep convolutional neural network (CNN) [15]. A similar network was first introduced in [33] and successfully utilized to predict the formaldehyde concentration for the laboratory calibration measurements. For this contribution, the structure of this network was adapted to predict not only one gas concentration at a time, but the concentrations of all gases offered during calibration, i.e., acetone, ethanol, formaldehyde, toluene, the total concentration of all VOCs (VOC_{sum}), and also the inorganic gases carbon monoxide and hydrogen. The CNN structure was derived from the original ResNet model from [33] to reduce the overall complexity.

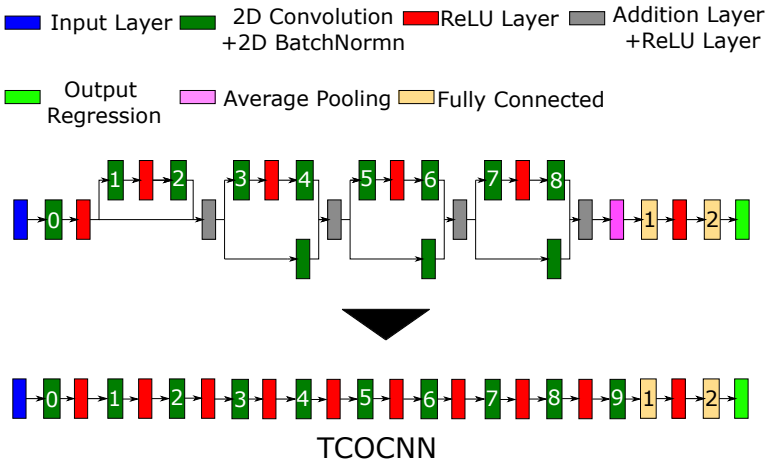


Figure 3. Original network structure from [33] together with the newly derived general architecture of the TCOCNN.

To build a gas-specific model, the general architecture as illustrated in Figure 3 was used. For each gas, the data were randomly split according to the gas mixtures into 70% training data, 10% validation data, and 20% for testing. After the data split, a neural architecture search (NAS) was performed on the training and validation data. This approach searches through a predefined search space of the parameters of the neural network (Table 3) to find the optimal hyperparameters for each gas concentration to be predicted by minimizing the RMSE for the validation data [16].

The NAS varies the parameters listed in Table 3 using a Bayesian optimization search with the remaining parameters, as specified in Table 4. In total, 30 different combinations of the parameter were tested, and the model with the smallest validation RMSE was considered the best model. The Bayesian optimization strategy was chosen to speed up the NAS. Since the training process of one TCOCNN on a GPU already requires up to 20 min, an extensive search through the complete search space would have not been feasible. Therefore, Bayesian optimization was performed to find an acceptable solution in a reasonable time. The optimization that was performed in this specific case was based on the Gaussian process method [34], and the optimized cost function was the validation RMSE.

Table 3. Parameter ranges for every neural architecture search (NAS).

Initial Learning Rate (Log Scale)	Number of Filters (First Two Layers)	Kernel Size (First Two Layers)	Stride Size (First Layer)	Dropout	Number of Neurons (FC)
1×10^{-4}–9×10^{-3}	60–240	40–80	15–45	30–50%	1000–2500

Table 4. Parameters, which are kept constant during the evaluation.

Parameter	Value
L2-regularization	0.0001
Stride size, even layer	1×2
Kernel size, other layer	1×2
Learn rate drop rate	0.9
Mini-batch size	50
Stride size, odd layer	1×1
Epochs	75
Learn rate drop period	2

The parameters chosen for this optimization were the initial learning rate, the number of filters in the first two layers, the kernel size of the first two layers, the stride size of the first layer, the dropout rate, and the number of neurons in the last fully connected layer [15]. The initial learning rate had to be a part of the optimization as this parameter should be adjusted according to the network complexity. The hyperparameters of the first convolutional layers have proven to have a large influence on the prediction quality and were therefore an important part of the optimization. Additionally, the dropout rate and the number of neurons in the fully connected layers are also parameters worth considering. The ranges for the different target gases were based on the best parameters found in [33]. For carbon monoxide and ethanol, the NAS had to train 60 different TCOCNNs to reach sufficient results as the model building seemed to be more difficult for these gases. Furthermore, the NAS for ethanol had to be restricted to a range from 15 to 35 for the stride size of the first layer to find a suitable result faster.

2.3. Data Evaluation

As a carcinogenic gas, formaldehyde is of great importance for indoor air quality. Thus, as the first step, we evaluated the suitability of the model for predicting the formaldehyde concentration in the ppb range in a complex mixture of other gases [33]. Here, a model for formaldehyde was trained on the initial calibration dataset with a gas-mixture-based data split of 70% training, 10% validation, and 20% for testing.

To determine the required complexity of the calibration (note that one hundred unique gas mixtures offered for 20 min each resulted in a total calibration time of 33 h; the extended calibration with higher VOC concentrations, therefore, required almost 7 d in total), the same model was built with fewer core samples and/or fewer unique gas mixtures to reproduce the results achieved in [33] and also to show the influence of more core samples compared to more unique gas mixtures (UGM).

In the next step, the effect of sensor drift, which is often observed for chemical and especially MOS gas sensors [35,36], was examined. Here, three different models were compared for the prediction of the formaldehyde concentration. For the first model, only the initial calibration including extended concentrations was used for model building and the data of the second part of the first recalibration (with extended concentrations) were used for testing. This model should show significant sensitivity to sensor drift and various effects on the prediction quality such as offset or linearity errors and increased uncertainty. The second model then uses only three instead of all four gas-sensitive layers from the sensor system. The excluded layer is the one most prone to sensor drift as observed in

previous studies [37], and the model trained on these data was used to investigate the cause of the different drift effects. The last model then includes all gas-sensitive layers for training, but extends the calibration data to include the 100 unique gas mixtures of the first recalibration in the standard concentration range. By including parts of the first recalibration in the training, it was expected that the model could suppress drift effects, as these were included in the training data [38]. Again, the second part of the first lab recalibration was used to test the prediction of these models.

After validating the TCOCNN model in general, the results achieved for the laboratory tests with the deep-learning approach were compared with the FESR model published previously [13]. Therefore, TCOCNN models for all seven targets were trained with the help of the NAS on the initial calibration and the first part of the first recalibration to reduce the drift effects. The data split was performed as explained before (70% training, 10% validation, and 20% testing). For comparison, the RMSEs on the test data of the different models are compared for the different gases. This step allows comparing the prediction quality and capability of the different data-driven models.

After demonstrating the quality of the prediction of the TCOCNN approach for the lab data, the deep network was also applied to data from real indoor air environments during field tests. Again, the models were trained using the lab calibration data with the complete initial calibration and the first part of the first recalibration to predict all trained targets during the field test. First, the overall prediction of the TCOCNN for the field test data was compared to the FESR model. This should indicate if the predictions of the FESR model and the TCOCNN are consistent. Furthermore, the standard deviations of both predictions were calculated to estimate the uncertainty of a prediction based on one temperature cycle assuming that the gas concentration was changing only slowly during the field test. To determine the standard deviation, a period with minimal signal changes was selected, here between 4 October 12:00, and 5 October 0:00, and the model predictions during this period were smoothed with the help of a sliding window with a length of 1 h. The standard deviation between the original model output and the smoothed data was calculated as an estimate of the noise level of the different models.

Next, the predictions obtained during release tests were compared to investigate the quantitative performance of the two different models. Furthermore, the prediction qualities of the TCOCNN models were analyzed regarding their cross-influence. In addition, the TCOCNN output was compared with the results obtained from the analytical instruments to further evaluate the capabilities compared to state-of-the-art systems.

Finally, the models were tested regarding their capability to detect gases not contained in the calibration, but belonging to the same chemical class as one trained gas (Table 5). Here, release tests performed with m/p-xylene (an aromatic) and isopropyl alcohol (an alcohol) were considered to determine the ability of the trained models to extrapolate to similar chemical compounds. This would show if the systematic approach with the MOS sensor, dynamic operation, and ML modeling could quantify individual gas components or provide an estimate of the total concentration of a certain chemical class.

Table 5. Chemical classes investigated in this publication [13].

Chemical Class (Representative)	P90 in µg/m^3 (ppb)	P95 in µg/m^3 (ppb)
Alcohols (ethanol)	320 (∼170)	520 (∼280)
Aldehydes (formaldehyde)	340 (∼270)	480 (∼390)
Aromatics (toluene)	190 (∼50)	370 (∼90)
Ketones (acetone)	250 (∼100)	420 (∼170)

3. Results

3.1. Calibration Results

Figure 4 shows that, in general, fewer data samples significantly increased the RMSE and also the uncertainty or rather variation of the RMSE. The specific RMSE mean and variance values illustrated in Figure 4 were based on the RMSEs achieved on the same training, validation, and testing data in 10 different runs using the TCOCNN.

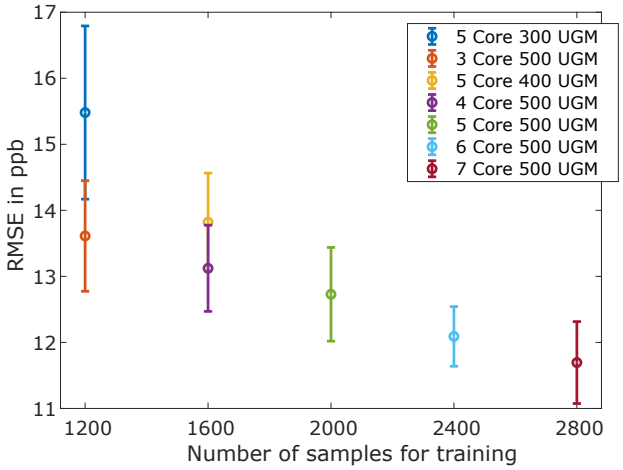


Figure 4. Obtained root-mean-squared error (RMSE) values for formaldehyde vs. the number of core samples and unique gas mixtures.

Obviously, fewer data samples degraded the prediction quality. Furthermore, Figure 4 illustrates with the comparison of five core samples for three-hundred UGM and three core samples for five-hundred UGM, i.e., both with one-thousand five-hundred cycles in total, that the number of unique gas mixtures is more important to achieve a high prediction quality than the number of core samples. Thus, a calibration should be biased towards containing more UGM with a shorter duration, resulting in fewer usable core samples, in accordance with the results of Robin et al. [33]. Note, however, that reducing the number of core samples did not decrease the overall duration of the calibration linearly, as temperature cycles recorded during a change of the gas composition cannot be evaluated. Moreover, Figure 4 illustrates that the difference between four core samples for five-hundred UGM and five core samples for four-hundred UGM, i.e., a total of two-thousand cycles each, resulted in only a minimal difference of the RMSE. Thus, it can be assumed that more than 500 UGM would not lead to a significant further reduction of the RMSE. This is also shown by the RMSE stagnating for more than five core samples, in agreement with previous results [33]. The best RMSE for formaldehyde for this dataset was achieved for five-hundred unique gas mixtures and seven core samples with an RMSE of 11.7 ppb. Nevertheless, because of the synchronization errors between the gas-mixing apparatus and the recording system, all further measurements were based on five core samples only, as these were always recorded under stable conditions.

For Figure 5a,b, the models were trained on the initial calibration data and the prediction was performed for the second part of the first recalibration (with extended gas concentrations). Figure 5c shows the results of a TCOCNN model that was trained on data including the initial calibration and the first part of the first recalibration with the prediction again performed on the second part of the first recalibration. Figure 5a shows that without any drift compensation, the prediction performance degraded severely over a period of six weeks (first field test period in a normal office environment) with a strong bias towards lower predicted concentrations and much higher uncertainty or variance of the prediction. Note that the TCOCNN did not predict negative concentration values; instead, many low concentrations were predicted as 0 ppb. Figure 5b illustrates that the one gas-sensitive layer that was excluded accounted for most of the drift, i.e., the major

part of the bias and the higher scatter of the predictions. Thus, excluding this layer already significantly improved the prediction quality, as indicated by the reduced variance, smaller linearity error, and reduced offset. The remaining linearity error can be attributed at least in part to a change of the formaldehyde test gas bottle between the initial calibration and the first recalibration. The test gas bottle concentrations had an uncertainty of 20%, which was most probably the cause for the remaining systematic linearity error. However, the gas-sensitive layer left out for the quantification of formaldehyde is in fact important for the detection and quantification of other VOCs, such as toluene, and also to reduce the cross-sensitivity to these gases. Thus, all available information should be used for building a comprehensive data-based model, and a different approach for reducing the drift is necessary. As previously reported, including data that already contain drift in the calibration, a so-called extended calibration [38], can improve the performance considerably, so this approach was also tested here. Figure 5c shows that the prediction based on extending the calibration dataset to include also data from the first recalibration after four weeks of field operation significantly improved the prediction quality. This model showed only a slight increase in the variance and a small offset and linearity error. Again, the linearity error could also be due to the change of the test gas bottles between the initial calibration and first recalibration, which would result in a systematic error.

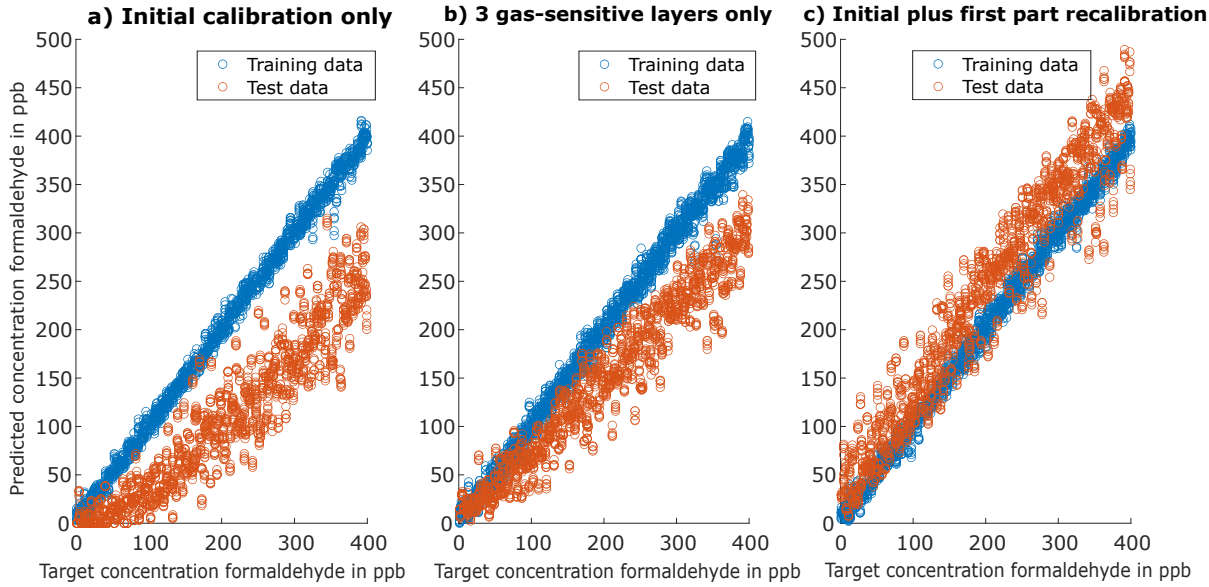


Figure 5. Evaluation of the target gas formaldehyde over several weeks. Test data always consist of the data of the second part of the first recalibration. (**a**) Results based on training with initial calibration only. (**b**) Results based on training with initial calibration only, but without one gas-sensitive layer that shows large drift over time. (**c**) Results based on training with the initial calibration and the first part of first recalibration.

This following section of the results demonstrates the performance achieved with the TCOCNN approach for the prediction of all gases present in the calibration dataset. The results obtained for formaldehyde on the number of core samples and unique gas mixtures, as well as the proven approach for compensating sensor drift were transferred to build the models for the other target gases. Thus, the presented results were always based on training with five core samples and extended calibration containing data from the initial calibration and the first part of the first recalibration. As before, a gas-mixture-based validation was performed. Accordingly, the data were split into 70% training, 10% validation, and 20% testing. In Table 6, the hyperparameters selected with the help of the training data, the validation data, and the NAS are listed. This shows that for most gases, the stride sizes of the first layer need to be larger than in the following layers.

Table 6. Optimized hyperparameters found during neural architecture search (NAS).

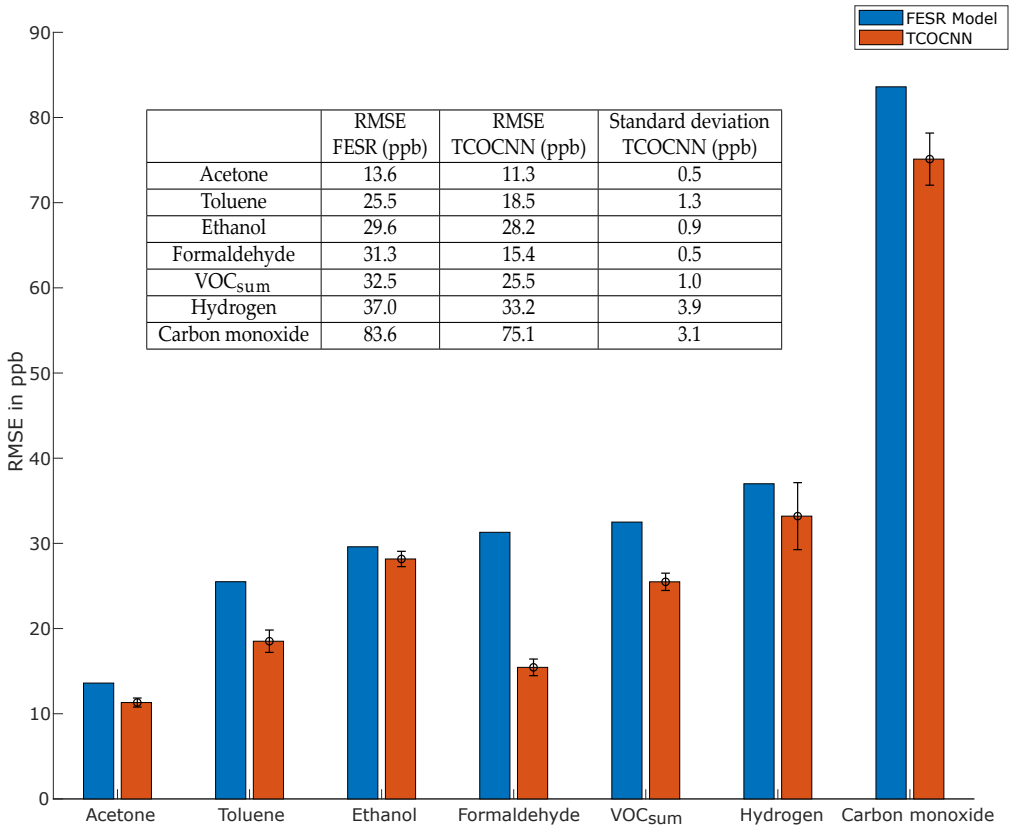
Substance	Initial Learning Rate (Log Scale)	Number of Filters (First Two Layers)	Kernel Size (First Two Layers)	Stride Size (First Layer)	Dropout	Number of Neurons (FC)
Acetone	1×10^{-4}	142	48	34	31.65%	1084
Toluene	2×10^{-4}	240	69	17	37.06%	2462
Formaldehyde	3×10^{-4}	183	55	18	49.75%	1188
Ethanol	2×10^{-4}	228	73	34	49.55%	2310
VOC_{sum}	1×10^{-4}	77	78	44	30.89%	1373
CO	6×10^{-4}	151	52	30	34.96%	2468
Hydrogen	1×10^{-4}	77	41	19	49.07%	1088

Figure 6 illustrates the RMSE results of the 70/10/20 data split on the initial calibration and the first part of the first recalibration for the FESR model (blue) and the mean value plus the standard deviation of the TCOCNN (orange) with the hyperparameters listed in Table 6. This shows that the TCOCNN achieved at least the same and often a significantly lower RMSE compared to the FESR model for all gases. The most significant improvement was achieved for formaldehyde, where the mean RMSE of the TCOCNN was less than half of the RMSE of the FESR model (15.4 ppb for TCOCNN vs. 31.3 ppb for FESR). This significant improvement for formaldehyde was probably due to the fact that the underlying model of the TCOCNN was originally optimized for formaldehyde quantification, i.e., more hyperparameters were optimized for formaldehyde than for the other gases. Thus, extending the NAS to include also the parameters that were kept constant in this study might result in similar improvements also for the other gases. Nevertheless, the results clearly showed that the TCOCNN models outperformed the FESR models regardless of the gas on which they were trained. Moreover, the variations of the RMSE caused by the different initializations of the TCOCNN were relatively small; thus, a stable model was achieved even if the network was trained only once. The variance of the RMSE is not given for the FESR method as the PLSR is deterministic, i.e., always produces the same result with the hyperparameters specified during the 10-fold cross-validation.

3.2. General Field Test Results

After showing that the TCOCNN can successfully predict the concentration of various gases in complex laboratory environments, this part focuses on quantifying the trained gases in a real indoor air environment. First, the general prediction quality of the TCOCNN for the various gases was compared to the predictions of the FESR model. Figure 7 illustrates the prediction of indoor air between September 26 and October 18 for formaldehyde and hydrogen. These two gases were chosen as they showed relevant aspects; the results for the other gases were similar. First, a constant offset was observed between both models, with the FESR model for formaldehyde indicating significantly higher concentrations than the TCOCNN model (average offset 140 ppb), while for hydrogen, the prediction of the TCOCNN was slightly higher with an average offset of 98 ppb. These differences were probably caused by the presence of additional gases in the room, which were not part of the calibration and were therefore not (fully) compensated by the data-based models and/or by the gas concentrations of the trained gases in the indoor environment outside of the trained ranges. Without reference measurements, it is not possible to determine which value is correct; in fact, both could be similarly off with one prediction being too high, the other too low. Nevertheless, at least for hydrogen, the baseline of the TCOCNN model seemed to be more realistic, as after ventilation events, the TCOCNN model indicated concentrations around 500 ppb, corresponding to the natural background level [39]. For formaldehyde as well, the lower average concentrations indicated by the TCOCNN model seemed more realistic, as the FESR model indicated concentrations well above the

WHO recommended limit value of 80 ppb [40]. This will be investigated in the future with formaldehyde reference measurements as described in the relevant standards [40]. More importantly, however, both models were in agreement concerning the relative changes of the gas concentrations, which would be required to indicate changes in the indoor air quality, e.g., for demand-controlled ventilation.



	RMSE FESR (ppb)	RMSE TCOCNN (ppb)	Standard deviation TCOCNN (ppb)
Acetone	13.6	11.3	0.5
Toluene	25.5	18.5	1.3
Ethanol	29.6	28.2	0.9
Formaldehyde	31.3	15.4	0.5
VOC_{sum}	32.5	25.5	1.0
Hydrogen	37.0	33.2	3.9
Carbon monoxide	83.6	75.1	3.1

Figure 6. Comparison of the RMSE values obtained with FESR (adapted from [13]) and the TCOCNN.

In addition, the FESR models showed much higher noise or short-term fluctuations compared to the TCOCNN models. This was specifically dominant for hydrogen and formaldehyde. To quantify this effect further, the standard deviation was calculated for all gases for both models. For this, a fairly stable time period without release events (4 October 12:00, to 5 October 00:00) was chosen, and the standard deviation between the model predictions and their hourly average as the estimated mean signal was calculated, resulting in the values given in Table 7. The ratio of the noise levels of the FESR model vs. the TCOCNN model was between 1.4 (for ethanol) and 5.2 (for hydrogen) to 6.2 (for formaldehyde), which is also evident from Figure 7. Furthermore, some predictions of the FESR model were below zero, which was not the case for the TCOCNN. These short events were caused by ventilating the room in which the experiments were performed, which probably resulted in very low gas concentrations below the calibrated range. Thus, it was not surprising that the models were not able to quantify the gases correctly in these conditions. Taking all observations into account, we concluded that for all gases, the overall quality of the TCOCNN model was more suitable for monitoring real indoor air as no false-negative values were obtained and the noise in a room where the gas composition changes only slowly is much lower. The absolute error of both models can only be determined with calibrated reference measurements, which were not available for this study, but which will be performed in the near future.

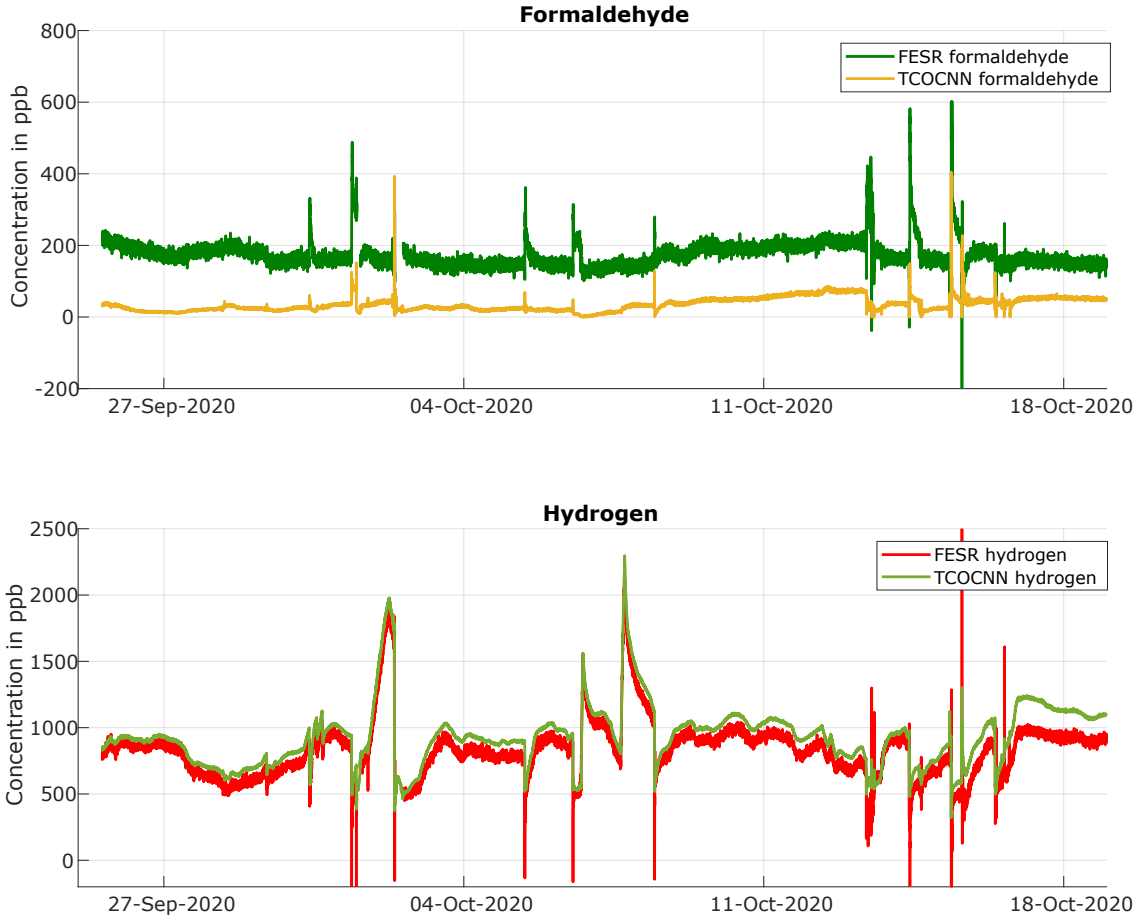


Figure 7. Comparison of the results obtained during field tests with the FESR and TCOCNN models for formaldehyde and hydrogen.

Table 7. Standard deviation during the field test of the TCOCNN and FESR.

	Standard Deviation FESR (ppb)	Standard Deviation TCOCNN (ppb)
Toluene	9.7	2.0
Formaldehyde	9.9	1.6
Carbon monoxide	33.6	5.5
VOC_{sum}	15.8	4.5
Acetone	5.7	2.6
Ethanol	15.0	10.7
Hydrogen	19.0	3.7

3.3. Results of the Release Tests for the Trained Gases

After the general observations of the similarities and differences between both models, this section focuses on the release tests performed during the field test period. Since formaldehyde as a carcinogenic gas and carbon monoxide as a toxic gas could not be actively released, the following results concentrated on release tests of acetone, ethanol, toluene, and hydrogen.

First, the releases of individual VOCs were analyzed in detail. Figure 8a illustrates the release of toluene on November 2 (Test 7). Both the FESR and the TCOCNN models

predicted similar average concentrations over time, but with higher noise for the FESR model, as observed before. The FESR model indicated an increase of approximately 600 ppb, which is in accordance with the amount released, cf. (Table 2), while the TCOCNN indicated a slightly smaller increase of 500 ppb. In addition to the MOS gas sensors, the release was also monitored with the portable GC-PID (X-pid 9500), which indicated a similar rapid increase and slow decrease of the toluene concentration, but a higher absolute value with an increase of approximately 700 ppb. Note, however, that the limit of quantification (LOQ) for the X-pid 9500 is 1 ppm for toluene according to the manufacturer (500 ppb for acetone).

For acetone, Figure 8b shows an offset between the absolute concentrations predicted by the TCOCNN and FESR models: the baseline concentration of the TCOCNN model was approximately 72 ppb, while the FESR model indicated a baseline concentration of approximately 120 ppb. This difference of 48 ppb was also observed during the release of acetone with both models indicating the same increase of approximately 450 ppb and showing the same decline vs. time. The X-pid 9500 indicated a baseline value similar to the FESR model, but again predicted a larger increase of approximately 700 ppb with the same shape over time. The expected increase caused by the amount of acetone released was 600 ppb, but the actual concentration at the site of the sensors could be higher or lower depending on the distribution in the room and also secondary effects, such as adsorption on surfaces. Nevertheless, both data-driven models were clearly capable of detecting acetone with a high temporal resolution.

Figure 8c illustrates the release of hydrogen from a test gas bottle with an expected maximum concentration increase of approximately 2 ppm. The graphs show the corresponding values indicated by the TCOCNN and FESR models, as well as the GC-RCP reference instrument. As already observed in Figure 7, the baseline concentration indicated by the TCOCNN model was slightly higher compared to the FESR model, and the noise level of the TCOCNN was much smaller compared to the FESR model. The increase of the hydrogen concentration indicated by both models was similar (around 1500 ppb) and realistic compared to the amount of released gas, especially considering the relatively slow release over several hours, where some gas exchange and therefore loss of hydrogen is unavoidable. The GC-RCP (limit of detection 10 ppb) indicated a similar increase of the hydrogen concentration, but with an even lower baseline compared to both sensor models. We suspect that the RCP was underestimating the hydrogen concentration slightly [13], which was especially evident during ventilation events where the GC-RCP indicated concentrations well below the natural background concentration of 500 ppb [41]. Additionally, the RCP also showed a larger noise level compared to the TCOCNN model. Regarding the other models, only ethanol and carbon monoxide showed a cross-influence, which was relatively small compared to the released amount of hydrogen (see Figure A1).

Finally, Figure 8d illustrates the values of the respective TCOCNN models during a simultaneous release of acetone, ethanol, and toluene. For toluene, the peak increase was approximately 400 ppb, which is slightly lower than during the release test of toluene only, similarly for acetone with an increase of approximately 360 ppb. Nevertheless, all three models detected the release of the various compounds with a high temporal resolution, which is especially evident when observing the different shapes of the release peaks: acetone with the lowest boiling point showed the sharpest peak, while toluene with a comparatively high boiling point showed a much broader and rounded release peak. To further elucidate the simultaneous evaluation of the various data-based models, Figure A1 shows the behavior of all other models during those release tests, and the following figure shows as an example all calculated model outputs during a specific release test including the VOC_{sum} model, indicating the sum concentration of all VOCs.

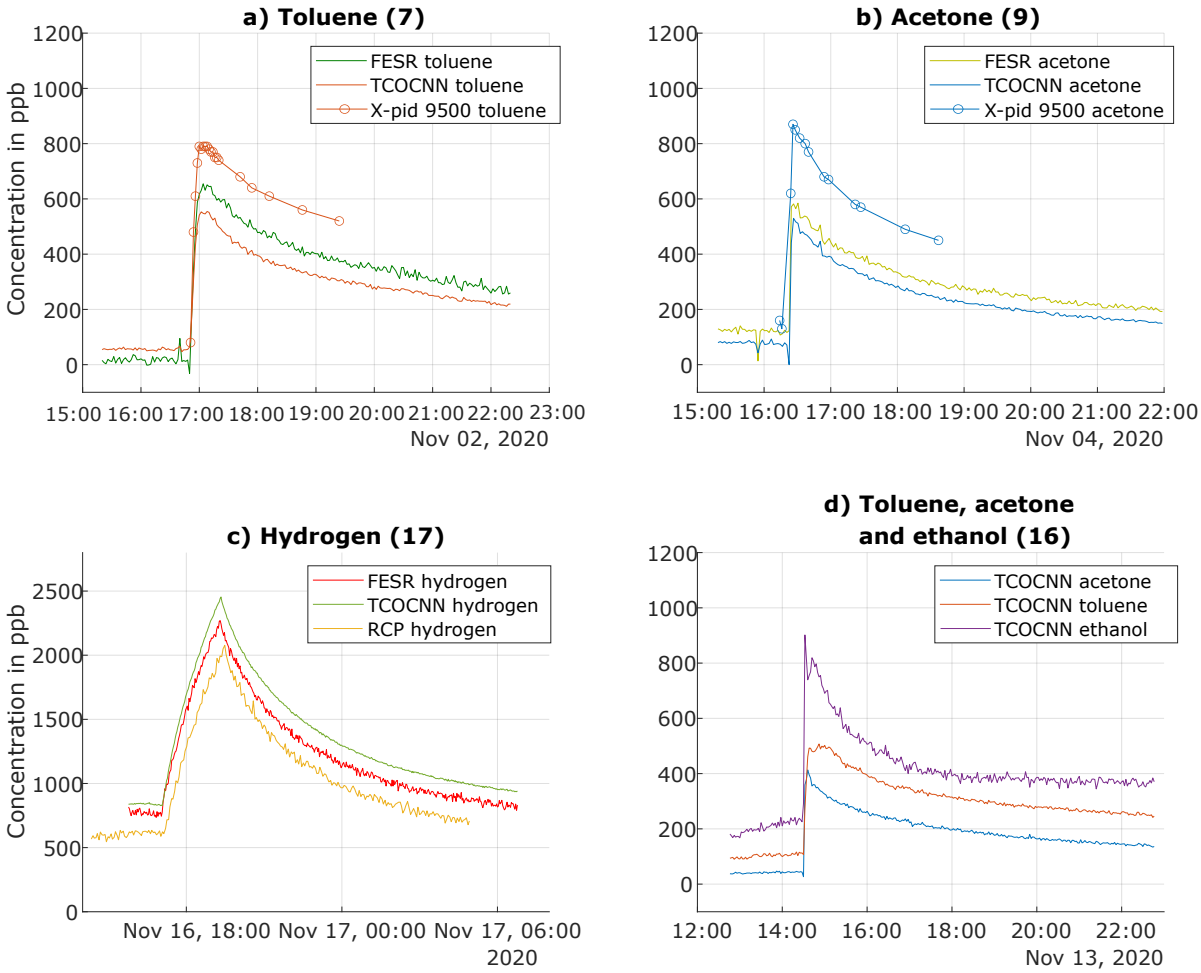


Figure 8. Prediction of gas concentrations during release tests for various trained gases using different evaluation models and a comparison with the results of the analytical instruments (adapted from [13]).

Figure 9 illustrates the TCOCNN model output during two simultaneous releases of acetone and toluene. The increase of the acetone and toluene concentrations indicated by the models was similar to the previous test shown above (Tests 7 and 9). For toluene and acetone, the signal increased by approximately 550 ppb, which is close to the expected value of 600 ppb. For both gases, the X-pid 9500 again indicated a significantly higher concentration increase, but with a similar shape over time. Parallel TD-GC-MS (LOQ approximately 50 ppb for toluene) analysis with samples taken over 30 min intervals showed absolute concentrations, as well as an increase of the toluene concentration of approximately 550 ppb, which is very similar to the values obtained from the TCOCNN model; acetone was not evaluated with the TD-GC-MS method in this study as the sampling protocol would have to be adjusted for accurate quantification of this VVOC. Again, the FESR model showed an offset compared to the TCOCNN model with slightly higher baselines and also somewhat larger concentration increases for both gases; the absolute values of the FESR model were between the results obtained with the X-pid 9500 and the TD-GC-MS. Note, however, that the GC-MS was not calibrated before these measurements. Finally, both the FESR and TCOCNN models, as well as the X-pid 9500 indicated that the toluene increase during the second simultaneous release was much slower (no Tenax samples were collected during this second release). This slower increase was probably caused by the significantly lower temperature during the second release (at night), resulting in much slower evaporation of toluene. Again, all three methods—X-pid 9500 and both

MOS data-based models—showed the same shape over time, indicating their potential for monitoring IAQ in real-time.

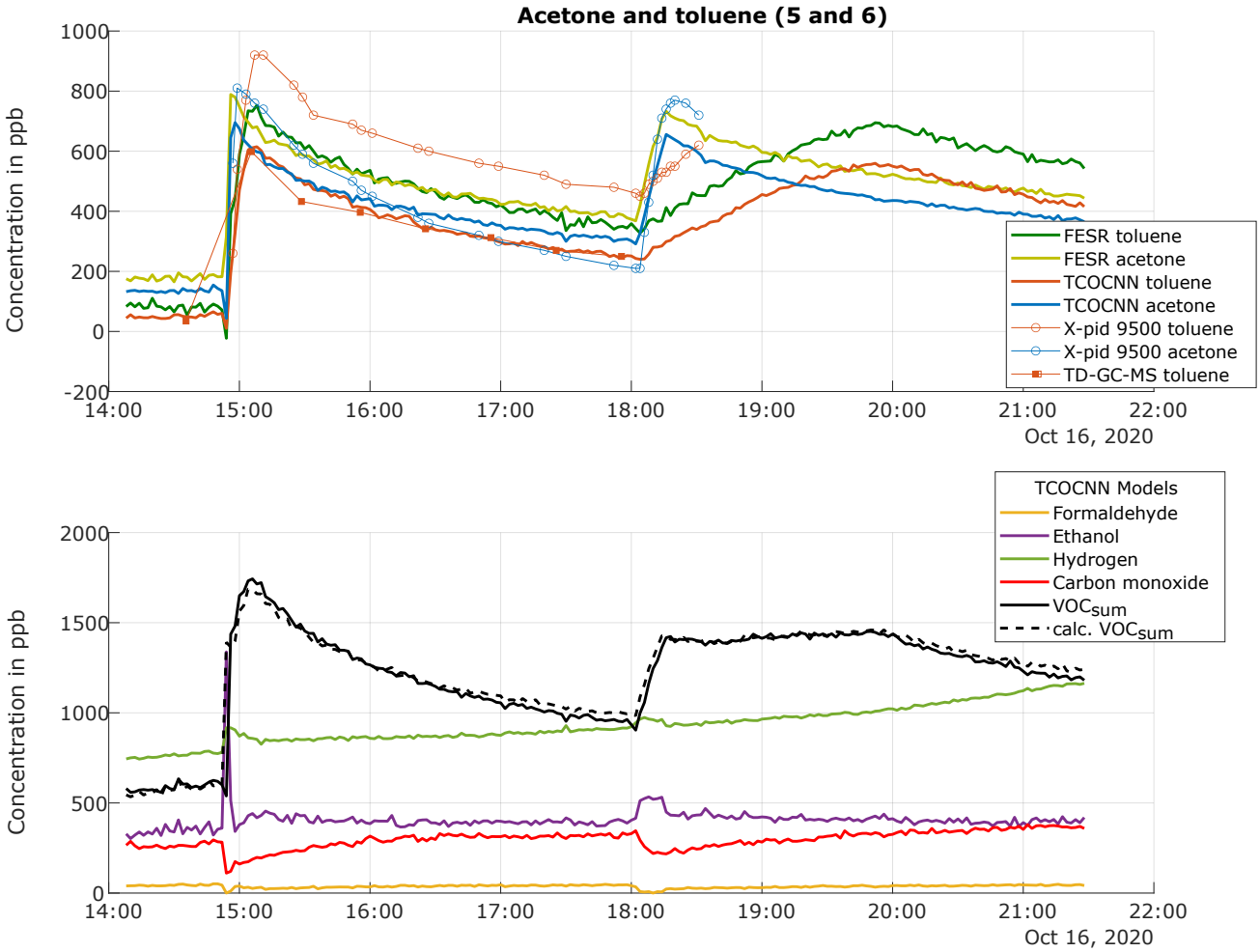


Figure 9. Prediction of gas concentrations during Release Tests 5 and 6 (acetone and toluene) showing the various models trained compared to the analytical measurements (adapted from [13]).

The outputs of the other TCOCNN models are shown in the lower part of Figure 9. Note that the trained VOC_{sum} model (solid black line) actually showed a nearly identical behavior compared to adding the concentrations indicated by all four separate VOC models for acetone, ethanol, formaldehyde, and toluene with an increase of approximately 1100 ppb during the first combined release. The VOC_{sum} model also showed the different evaporation speeds during the second release with a first fast increase caused by the release of acetone followed by an almost constant concentration due to the offsetting effects of increasing toluene and decreasing acetone concentrations.

For the other two VOCs, ethanol and formaldehyde, which were not released and were therefore expected to have a constant concentration, the TCOCNN models showed very little cross-sensitivity: a small drop was observed for formaldehyde and a short sharp increase for ethanol at the first release, but both models recovered their previous baseline quickly. These short-term effects were probably caused by the person performing the release test entering the room. However, a significant cross-sensitivity was observed for carbon monoxide, where the model output dropped by approximately 170 ppb during both releases and then recovered only slowly. A similar, but opposite effect was observed for the hydrogen model, which showed a minor increase during both release tests. Furthermore, a general baseline drift of the indicated hydrogen concentration was observed with a different

behavior over time compared to the released VOCs. This effect can be attributed to VOC decomposition, leading to an increase of hydrogen; in fact, the model predictions for hydrogen over time were in good agreement with the GC-RCP reference instrument [41].

The behavior regarding the cross-influence of different gases was observed for all release tests. Independent of the specific VOC released, the other VOC models showed only minor effects, while carbon monoxide showed a significant cross-sensitivity, which was probably caused by the comparatively low sensitivity of the SGP30 to carbon monoxide [37]. Hydrogen actually showed large variations during the field tests, which were not correlated with the release tests, indicating other sources inside the room with a diurnal pattern. The VOC_{sum} signal always accurately indicated the combined concentration of the various VOCs. The release of hydrogen did not result in a significant response of any other model, illustrating the high selectivity achieved for hydrogen, as previously reported for the FESR model [41].

3.4. Results of Release Tests for Gases Not Trained

To further elucidate the selectivity of the various models, release tests were performed with gases not included in the calibration, but from the same chemical classes, i.e., m/p-xylene as a second aromatic compound and isopropyl alcohol as a second alcohol. Again, we compared the performance of the TCOCNN model with the FESR model. Figure 10 illustrates the predictions of both models calibrated for toluene during the release of m/p-xylene (Test Number 14). Both the FESR model and the TCOCNN model indicated a similar evolution over time of the toluene concentration. The indicated increase for the FESR was approximately 450 ppb, so again, close to the theoretically expected increase of 600 ppb, while the increase with the TCOCNN was slightly smaller with 350 ppb. The X-pid 9500, on the other hand, showed a large offset with a baseline value of almost 500 ppb and an increase similar to the FESR model. These results showed that both data-driven models were capable of quantifying aromatics, i.e., chemicals from the same chemical class as the calibrated toluene, in agreement with previous results for VOC identification [42]. The other VOC models were not influenced by the release of m/p-xylene, indicating good selectivity (see Figure A2 in Appendix A) . The VOC_{sum} model also responded to the release of m/p-xylene, again similar to toluene. Finally, a significant cross-sensitivity of the carbon monoxide model was also observed during the release of m/p-xylene, similar to the release of toluene.

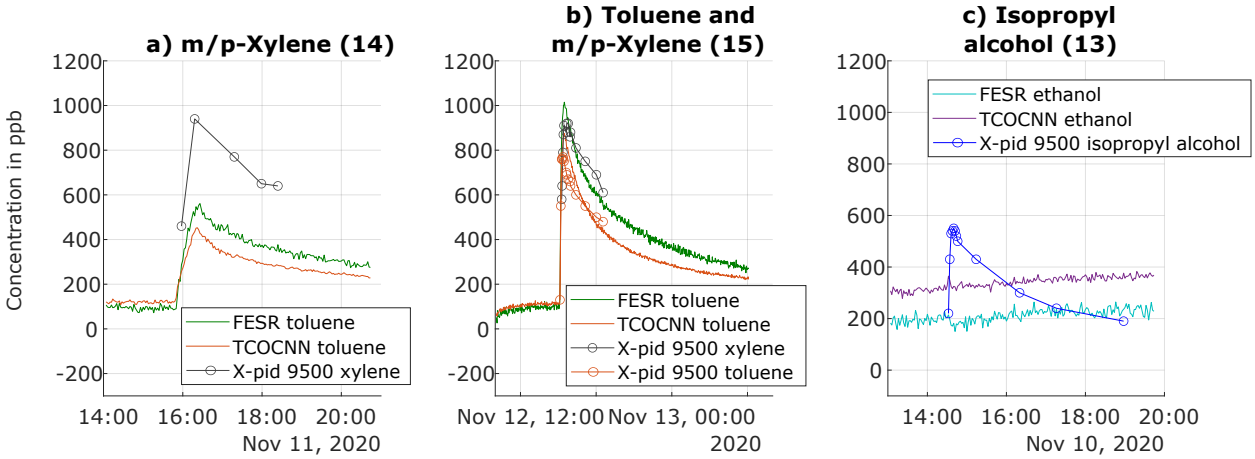


Figure 10. Prediction of gas concentrations for release tests of gases not contained in the calibration. (**a**) Release Test Number 14 of m/p-xylene. (**b**) Release Test Number 15 of toluene and m/p-xylene. (**c**) Release Test Number 13 of isopropyl alcohol (adapted from [13]).

The second release test in Figure 10 illustrates that the combined release of toluene and m/p-xylene could be detected by both data-based models, as well as the X-pid 9500, which can also discriminate between both aromatics, while the data-driven models could

not discriminate the two gases; the indicated increase was close to the sum of the previous releases. Again, a similar behavior was observed for the other TCOCNN models of the other target gases (see Figure A2 in Appendix A). Note that discrimination between both aromatic compounds might be possible with the data-based models if both gases are included separately in the calibration.

On the other hand, both the TCOCNN, as well as the FESR model did not indicate the release of isopropyl alcohol as ethanol, as shown in Figure 10c. Only a small upwards drift for ethanol is visible. We cannot provide a conclusive answer, but the upward drift already started before the release of isopropyl alcohol, so this was probably due to some unrelated background event. Interestingly, the FESR model showed a slight decrease when isopropyl alcohol was released, but otherwise, the same trend as the TCOCNN model. Regarding the other TCOCNN models, no model showed a significant reaction to this gas (see Figure A2 in Appendix A) not included in the calibration. Only the X-pid 9500 was capable of detecting and quantifying isopropyl alcohol as well. This result showed that the models were not always capable of indicating gases of the same chemical class, in this case alcohols. While this might be seen as a drawback, because a more complex calibration would be required to obtain a valid VOC_{sum} model, when considering formaldehyde, a high selectivity also to other aldehydes such as acetaldehyde is preferable to discriminate the health impact of the various gases.

4. Discussion

This contribution discussed the application of a deep-learning model for evaluating the complex data patterns recorded with a metal oxide gas multisensor (SGP30) in temperature-cycled operation (TCO) to independently determine multiple gas concentrations from a single sensor element. The novel deep-learning approach named the TCOCNN was based on a 10-layer CNN design. In this work, the performance of the TCOCNN was studied both concerning the optimization using different data configurations by varying the number of temperature cycles per gas mixture (core samples) and the number of unique gas mixtures (UGM). As expected, the number of UGM is more important to achieve a low RMSE of the prediction than the number of (basically redundant) samples. Note that, when taking the number of independent variables into account (six gases plus RH), the number of unique gas exposures for reliably achieving a stable ML model is actually fairly low. A full factorial calibration with seven parameters at four levels each would result in a total of more than 16,000 tests. Second, we studied the potential to suppress drift, which is often observed for MOS gas sensors, by extending the calibration data to also include data from an additional one-hundred gas exposures obtained in the second calibration run after four weeks of operation in the field. This extended calibration succeeded in greatly reducing the offset, linearity, and noise error observed when only the original calibration data were used to build the model. With this approach, a stable prediction of the formaldehyde concentration with an uncertainty of approximately 42 ppb was achieved over a period of at least 6 wk. Third, the TCOCNN approach was successfully tested not only for formaldehyde, but also for predicting at the ppb level the concentrations of the other gases included in the calibration plus an additional model for the sum of all VOCs, VOC_{sum}. Here, a neural architecture search with Bayesian optimization was performed to select suitable hyperparameters for the TCOCNN models. The results showed that stable models were reproducibly obtained with this approach, achieving a performance at least as good as the previous linear models in terms of the RMSE for the calibration data. The improvement was most significant for formaldehyde, where the RMSE was more than halved. The different initializations of the TCOCNN only resulted in negligible variation of the RMSE between 1 ppb and 8 ppb, indicating the excellent reproducibility of the model-building approach. Applying these models to data from the field tests showed that the TCOCNN models had lower noise for real field data compared to the previous FESR model and did not predict negative gas concentrations even when operated outside the calibrated gas concentration range. We did observe significant offsets between both ML models,

which were probably caused by unknown gases not contained in the calibration, but present during the field tests. This will require further analysis with calibrated reference instruments to determine which model provides higher absolute accuracy. However, variations of the gas concentrations can be accurately monitored with high temporal resolution, as demonstrated with various release tests during the field test period. In fact, the indicated concentration increases of the released gas closely matched the expected theoretical values and often significantly outperformed the mobile GC-PID and GC-RCP instruments used for comparison both when releasing hydrogen or single VOCs and when simultaneously releasing VOC mixtures. The best absolute agreement was observed for the TCOCNN model and the gold standard TD-GC-MS for toluene monitoring. Minimal cross-sensitivity was observed for the six gases tested in this study with only the carbon monoxide model showing a slightly higher cross-sensitivity to VOCs, probably due to the low overall sensitivity of the SGP30 to CO. Finally, release tests were performed with gases not contained in the calibration. Here, two different results were observed for m/p-xylene and isopropyl alcohol. While the TCOCNN for toluene was also able to detect and quantify m/p-xylene with reasonable accuracy, neither the ethanol model nor any other reacted to isopropyl alcohol. This shows that in some cases (toluene and m/p-xylene), the sensor actually detects a certain chemical class, here aromatics, while in others, the gases (ethanol and isopropyl alcohol), although belonging to the same chemical group, here alcohols, induce unique sensor response patterns allowing discrimination and quantification of the individual components. This aspect will require further examination, as both effects are beneficial in some ways and undesirable in others. Quantifying all gases from the same chemical group after the calibration of only a single representative would greatly reduce the complexity of the sensor calibration. On the other hand, being able to quantify individual gases even against others from the same chemical class is important for the accurate determination of relevant indoor pollutants such as formaldehyde (vs. acetaldehyde and others) or benzene (vs. toluene and xylene). The presented systematic approach could provide the basis for the development of high-performance application-specific VOC sensor systems taking target and interfering gases into account.

Regarding the computation time for hyperparameter tuning and model training, it should be noted that the TCOCNN model training requires much more time. While the FESR method requires up to 24 h for a full evaluation including hyperparameter optimization, the TCOCNN including NAS requires several days. Therefore, a reduction of the computational complexity of the TCOCNN is desirable for future investigations.

5. Conclusions and Outlook

All-in-all, the novel TCOCNN model presented here outperformed state-of-the-art ML models such as the FESR approach both in laboratory measurements and field tests, achieving higher accuracy and lower noise with the same temporal resolution, especially in real application environments. Furthermore, the TCOCNN model achieved similar quantification performance as the tested analytical systems, which however were more robust in the case of unknown gases.

On the other hand, the TCOCNN approach is still not fully investigated, as it is not yet clear on which features the model is basing its decision and how the hyperparameters influence the model performance for various target gases. Furthermore, the absolute accuracy has to be determined with calibrated reference instruments, which were not available for this study. Similarly, the selectivity and quantification performance for gases from the same chemical group as one of the trained gases needs to be studied further to make full use of this effect to reduce the calibration complexity while still achieving the required level of selectivity. Finally, we are planning to investigate methods such as transfer learning to reduce or even eliminate the required recalibration for drift compensation.

Author Contributions: Conceptualization, Y.R., T.B. and A.S.; methodology, Y.R., P.G. and J.A.; software, Y.R. and P.G.; validation, Y.R., T.B., P.G. and J.A.; formal analysis, Y.R.; investigation, Y.R.; resources, A.S.; data curation, J.A.; writing—original draft preparation, Y.R.; writing—review and editing, Y.R., T.B., C.S., J.A., T.S. and A.S.; visualization, Y.R.; supervision, Y.R., T.B., T.S. and A.S.; project administration, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this research was performed within the project "SE-ProEng" funded by the European Regional Development Fund (ERDF). We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and Saarland University within the funding program Open Access Publishing.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The underlying data are available on Zenodo and were published with our first publication on the field tests: DOI:10.5281/zenodo.4593853. Title: Measuring Hydrogen in Indoor Air with a Selective Metal Oxide Semiconductor Sensor: Dataset. Authors: Johannes Amann, Tobias Baur, Caroline Schultealbert, https://zenodo.org/record/4593853 (accessed on 17 May 2021).

Acknowledgments: We thank Rainer Lammertz Pure Gas Products for providing the Peak Performer 1 reference instrument and Dräger Safety AG & Co KGaA for providing the X-pid 9500 for this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
FE	Feature extraction
FESR	Feature extraction selection regression
FS	Feature selection
GC-PID	Gas chromatograph with photo-ionization detection
GC-RCP	Gas chromatograph with reducing compound photometer
IAQ	Indoor air quality
LOQ	Limit of quantification
MFC	Mass flow controller
ML	Machine learning
MOS	Metal oxide semiconductor
NAS	Neural architecture search
PLSR	Partial least squares regression
PM	Particulate matter
RH	Relative humidity
RFE	Recursive feature elimination
RMSE	Root-mean-squared error
SVOC	Semivolatile organic compounds
TCO	Temperature-cycled operation
TD-GC-MS	Thermo-desorption gas chromatography mass spectrometry
UGM	Unique gas mixtures
VOC	Volatile organic compounds
VVOC	Very volatile organic compounds

Appendix A

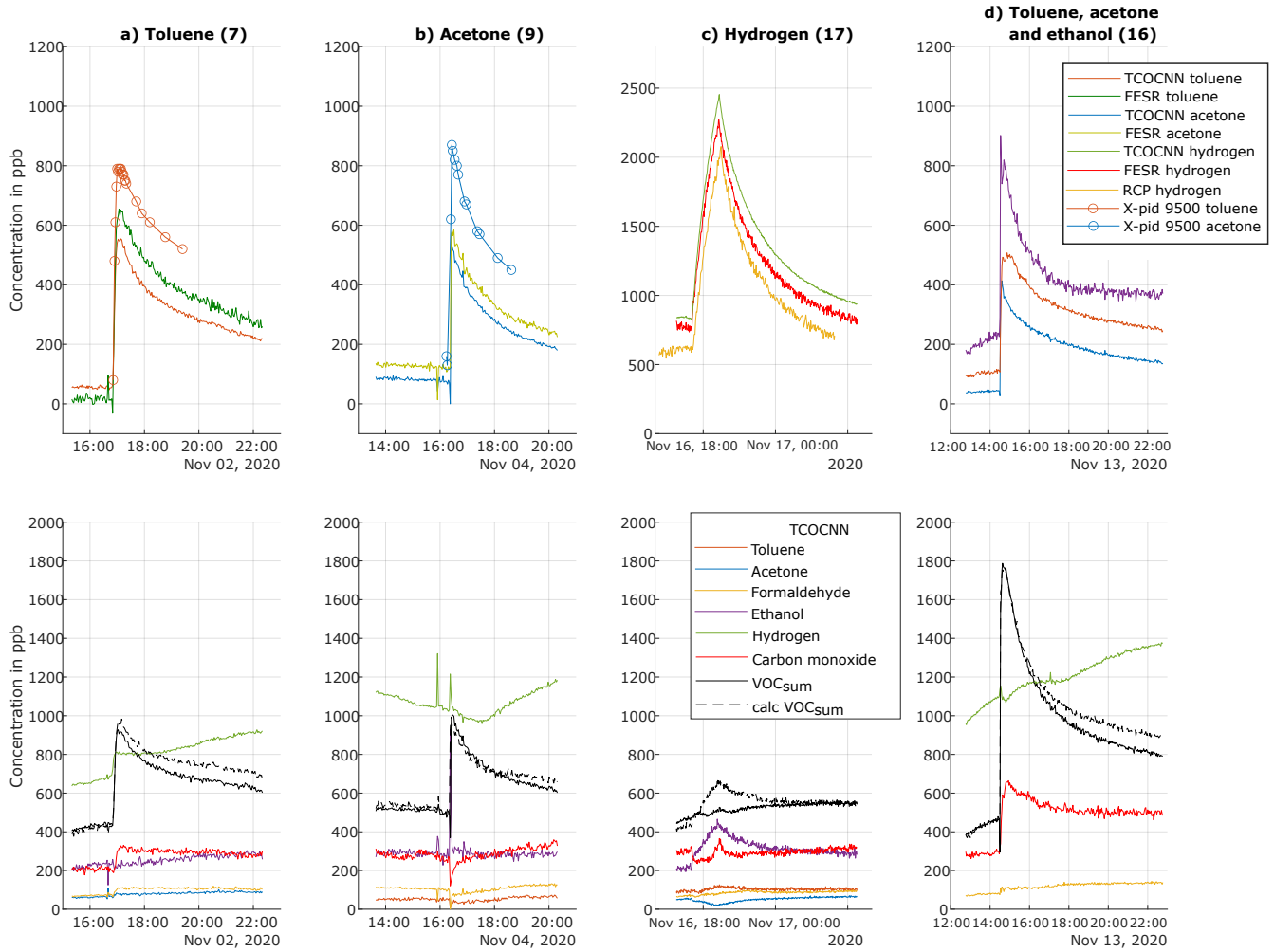


Figure A1. Prediction of gas concentrations during the release tests for various trained gases using different evaluation models and comparison with the results of analytical instruments together with all predictions of the TCOCNN for the other gases. (**a**) Release Test Number 7 of toluene. (**b**) Release Test Number 9 of acetone. (**c**) Release Test Number 17 of hydrogen. (**d**) Release Test Number 16 of toluene, acetone and ethanol (adapted from [13]).

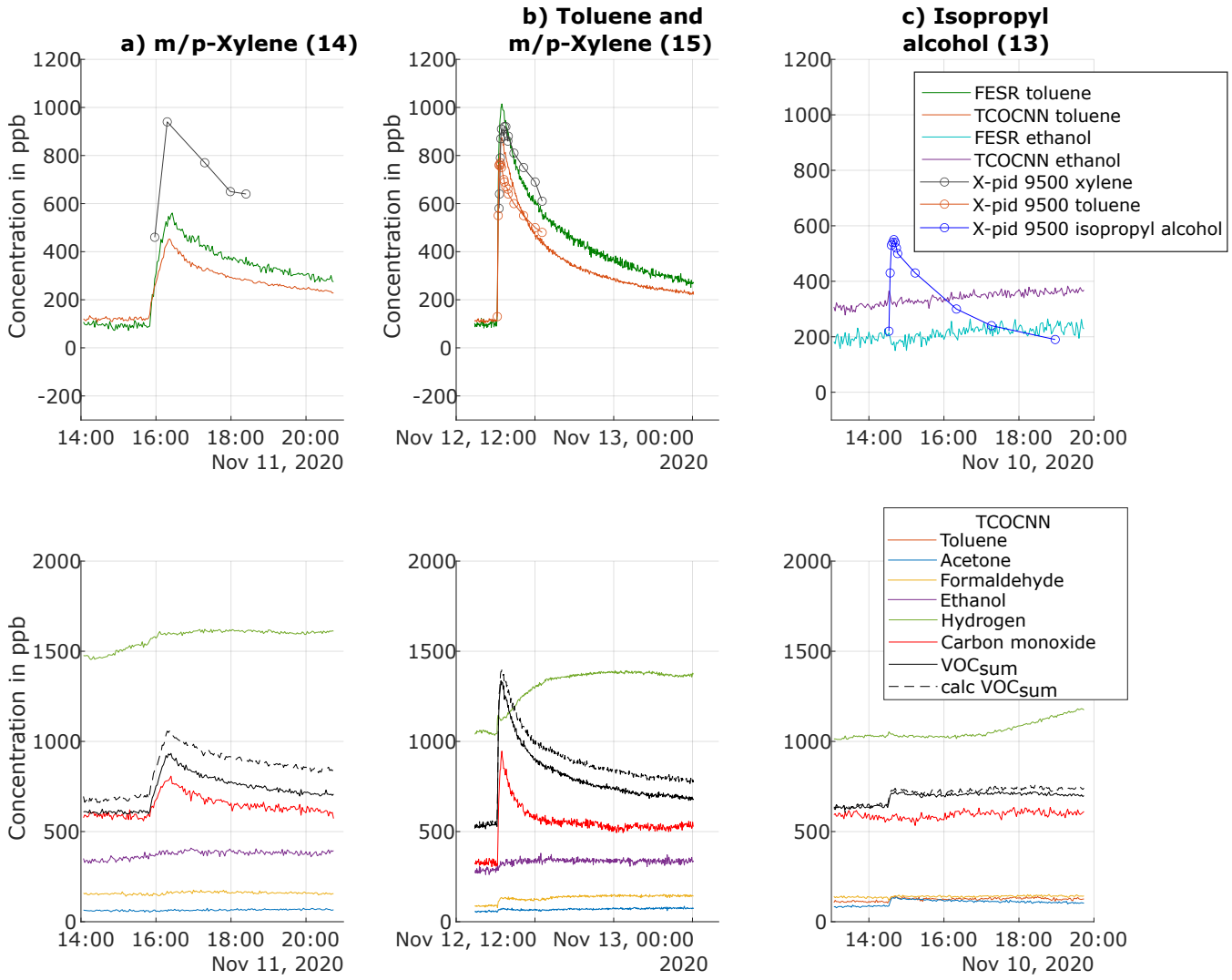


Figure A2. Prediction of gas concentrations for the release tests of gases not contained in the calibration together with all predictions of the TCOCNN for the other gases. (**a**) Release Test Number 14 of m/p-xylene. (**b**) Release Test Number 15 of toluene and m/p-xylene. (**c**) Release Test Number 13 of isopropyl alcohol (adapted from [13]).

References

1. Asikainen, A.; Carrer, P.; Kephalopoulos, S.; De Oliveira Fernandes, E.; Wargocki, P.; Hänninen, O. Reducing burden of disease from residential indoor air exposures in Europe (HEALTHVENT project). *Environ. Health* **2016**, *15*, 61–72. [CrossRef] [PubMed]
2. Hauptmann, M.; Lubin, J.H.; Stewart, P.A.; Hayes, R.B.; Blair, A. Mortality from Solid Cancers among Workers in Formaldehyde Industries. *Am. J. Epidemiol.* **2004**, *159*, 1117–1130. [CrossRef] [PubMed]
3. United Nations, Department of Economic and Social Affairs, Sustainable Development. Ensure Healthy Lives and Promote Well-Being for All at All Ages. Available online: https://sdgs.un.org/goals/goal3 (accessed on 15 October 2021).
4. Valero, E. *Advanced Nanomaterials for Inexpensive Gas Microsensors: Synthesis, Integration and Applications*; Elsevier: Amsterdam, The Netherlands, 2020.
5. Molhave, L.; Nielsen, G.D. Interpretation and Limitations of the Concept "Total Volatile Organic Compounds" (TVOC) as an Indicator of Human Responses to Exposures of Volatile Organic Compounds (VOC) in indoor air. *Indoor Air* **1992**, *2*, 65–77. [CrossRef]
6. Salthammer, T. Very volatile organic compounds: An understudied class of indoor air pollutants. *Indoor Air* **2014**, *26*, 25–38. [CrossRef]
7. Pettenkofer, M. *Über den Luftwechsel in Wohngebäuden*; Literarisch-Artistische Anstalt der J.G. Cotta'schen Buchhandlung: München, Germany, 1858.
8. Yeoman, A.M.; Shaw, M.; Carslaw, N.; Murrells, T.; Passant, N.; Lewis, A.C. Simplified speciation and atmospheric volatile organic compound emission rates from non-aerosol personal care products. *Indoor Air* **2020**, *30*, 459–472. doi: 10.1111/ina.12652. [CrossRef]

9. Coggon, M.M.; McDonald, B.C.; Vlasenko, A.; Veres, P.R.; Bernard, F.; Koss, A.R.; Yuan, B.; Gilman, J.B.; Peischl, J.; Aikin, K.C.; et al. Diurnal Variability and Emission Pattern of Decamethylcyclopentasiloxane (D5) from the Application of Personal Care Products in Two North American Cities. *Environ. Sci. Technol.* **2018**, *52*, 5610–5618. [CrossRef]
10. Mølhave, L. Indoor air pollution due to organic gases and vapours of solvents in building materials. *Environ. Int.* **1982**, *8*, 117–127. [CrossRef]
11. Schütze, A.; Baur, T.; Leidinger, M.; Reimringer, W.; Jung, R.; Conrad, T.; Sauerwald, T. Highly Sensitive and Selective VOC Sensor Systems Based on Semiconductor Gas Sensors: How to? *Environments* **2017**, *4*, 20. [CrossRef]
12. Haddad, S.; Synnefa, A.; Marcos, M.Á.P.; Paolini, R.; Delrue, S.; Prasad, D.; Santamouris, M. On the potential of demand-controlled ventilation system to enhance indoor air quality and thermal condition in Australian school classrooms. *Energy Build.* **2021**, *238*, 110838. [CrossRef]
13. Baur, T.; Amann, J.; Schultealbert, C.; Schütze, A. Field Study of Metal Oxide Semiconductor Gas Sensors in Temperature Cycled Operation for Selective VOC Monitoring in Indoor Air. *Atmosphere* **2021**, *12*, 647. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe Nevada, CA, USA, 3 December–6 December 2012; Curran Associates Inc.: Red Hook, NY, USA, 2012; Volume 1, pp. 1097–1105.
15. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, L.; Wang, G.; et al. Recent Advances in Convolutional Neural Networks. *arXiv* **2017**, arXiv:1512.07108v6.
16. White, C.; Neiswanger, W.; Savani, Y. BANANAS: Bayesian Optimization with Neural Architectures for Neural Architecture Search. *arXiv* **2020**, arXiv:1910.11858v3.
17. Vito, S.D.; Massera, E.; Piga, M.; Martinotto, L.; Francia, G.D. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sens. Actuators B Chem.* **2008**, *129*, 750–757. [CrossRef]
18. Szczurek, A.; Szecówka, P.; Licznerski, B. Application of sensor array and neural networks for quantification of organic solvent vapours in air. *Sens. Actuators B Chem.* **1999**, *58*, 427–432. [CrossRef]
19. Han, L.; Yu, C.; Xiao, K.; Zhao, X. A New Method of Mixed Gas Identification Based on a Convolutional Neural Network for Time Series Classification. *Sensors* **2019**, *19*, 1960. [CrossRef]
20. Wang, S.; Hu, Y.; Burgues, J.; Marco, S.; Liu, S.C. Prediction of Gas Concentration Using Gated Recurrent Neural Networks. In Proceedings of the 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Genova, Italy, 31 August–2 September 2020. [CrossRef]
21. Chen, Z.; Zheng, Y.; Chen, K.; Li, H.; Jian, J. Concentration Estimator of Mixed VOC Gases Using Sensor Array With Neural Networks and Decision Tree Learning. *IEEE Sens. J.* **2017**, *17*, 1884–1892. [CrossRef]
22. Xu, Y.; Meng, R.; Zhao, X. Research on a Gas Concentration Prediction Algorithm Based on Stacking. *Sensors* **2021**, *21*, 1597. [CrossRef]
23. Benrekia, F.; Attari, M.; Bouhedda, M. Gas Sensors Characterization and Multilayer Perceptron (MLP) Hardware Implementation for Gas Identification Using a Field Programmable Gate Array (FPGA). *Sensors* **2013**, *13*, 2967–2985. [CrossRef]
24. Feng, S.; Farha, F.; Li, Q.; Wan, Y.; Xu, Y.; Zhang, T.; Ning, H. Review on Smart Gas Sensing Technology. *Sensors* **2019**, *19*, 3760. [CrossRef]
25. Bastuck, M. Improving the Performance of Gas Sensor Systems with Advanced Data Evaluation, Operation, and Calibration Methods. Ph.D. Thesis, Department Systems Engineering, Shaker Verlag, Saarland University, Düren, Germany, 2019.
26. Rüffner, D.; Hoehne, F.; Bühler, J. New Digital Metal-Oxide (MOx) Sensor Platform. *Sensors* **2018**, *18*, 1052. [CrossRef]
27. Baur, T.; Bastuck, M.; Schultealbert, C.; Sauerwald, T.; Schütze, A. Random gas mixtures for efficient gas sensor calibration. *J. Sens. Sens. Syst.* **2020**, *9*, 411–424. [CrossRef]
28. Helwig, N.; Schüler, M.; Bur, C.; Schütze, A.; Sauerwald, T. Gas mixing apparatus for automated gas sensor characterization. *Meas. Sci. Technol.* **2014**, *25*, 055903. [CrossRef]
29. Schultealbert, C.; Baur, T.; Schütze, A.; Sauerwald, T. Facile Quantification and Identification Techniques for Reducing Gases over a Wide Concentration Range Using a MOS Sensor in Temperature-Cycled Operation. *Sensors* **2018**, *18*, 744. [CrossRef] [PubMed]
30. Baur, T.; Schütze, A.; Sauerwald, T. Optimierung des temperaturzyklischen Betriebs von Halbleitergassensoren (Optimization of temperature-cycled operation of semiconductor gas sensors). *TM-Tech. Mess.* **2015**, *82*, 187–195. [CrossRef]
31. Schultealbert, C.; Baur, T.; Schütze, A.; Böttcher, S.; Sauerwald, T. A novel approach towards calibrated measurement of trace gases using metal oxide semiconductor sensors. *Sens. Actuators B Chem.* **2017**, *239*, 390–396. [CrossRef]
32. Bastuck, M.; Baur, T.; Schütze, A. DAV³E a MATLAB toolbox for multivariate sensor data evaluation. *J. Sens. Sens. Syst.* **2018**, *7*, 489–506. [CrossRef]
33. Robin, Y.; Goodarzi, P.; Baur, T.; Schultealbert, C.; Schütze, A.; Schneider, T. Machine Learning based calibration time reduction for Gas Sensors in Temperature Cycled Operation. In Proceedings of the 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Glasgow, UK, 17–20 May 2021; pp. 1–6. [CrossRef]
34. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv* **2012**, arXiv:1206.2944v2.
35. Vergara, A.; Vembu, S.; Ayhan, T.; Ryan, M.A.; Homer, M.L.; Huerta, R. Chemical gas sensor drift compensation using classifier ensembles. *Sens. Actuators B Chem.* **2012**, *166–167*, 320–329. [CrossRef]

36. Artursson, T.; Eklöv, T.; Lundström, I.; Martensson, P.; Sjöström, M.; Holmberg, M. Drift correction for gas sensors using multivariate methods. *J. Chemom.* **2000**, *14*, 711–723. [CrossRef]
37. Amann, J.F. Möglichkeiten und Grenzen des Einsatzesvon Halbleitergassensoren im temperaturzyklischenBetrieb für die Messung der Innenraumluftqualität-Kalibrierung, Feldtest,Validierung. Master's Thesis, Universität des Saarlandes, Saarbrücken, Germany, 2021.
38. Bur, C.; Engel, M.; Horras, S.; Schütze, A. Drift compensation of virtual multisensor systems based on extended calibration. In Proceedings of the IMCS2014—The 15th International Meeting on Chemical Sensors (Poster Presentation), Buenos Aires, Argentina, 16–19 March 2014.
39. Schleyer, E.B.R.; Wallasch, M. *Das Luftmessnetz des Umweltbundesamtes*; Umweltbundesamt: Dessau-Roßlau, Germany, 2013.
40. WHO. *WHO Regional Office for Europe Centers of Disease Control, WHO Guidelines for Indoor Air Quality: Selected Pollutants*; World Health Organization: Copenhagen, Denmark, 2010; Volume 9, ISBN 978-92-890-0213-4.
41. Schultealbert, C.; Amann, J.; Baur, T.; Schütze, A. Measuring Hydrogen in Indoor Air with a Selective Metal Oxide Semiconductor Sensor. *Atmosphere* **2021**, *12*, 366. [CrossRef]
42. Schutze, A.; Gramm, A.; Ruhl, T. Identification of Organic Solvents by a Virtual Multisensor System With Hierarchical Classification. *IEEE Sens. J.* **2004**, *4*, 857–863. [CrossRef]