Chemometrics and Intelligent Laboratory Systems 237 (2023) 104808

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Chemistry-informed machine learning: Using chemical property features to improve gas classification performance

Yeram Kim [a], Chiehyeon Lim [a,b,*], Junghye Lee [c], Sungil Kim [a,b], Sewon Kim [d], Dong-Hwa Seo [e]

[a] Department of Industrial Engineering, UNIST, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, Republic of Korea
[b] Graduate School of Artificial Intelligence, UNIST, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, Republic of Korea
[c] Technology Management, Economics, and Policy Program, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea
[d] Taesung Environmental Research Institute, 56-20, Hoehak 3-gil, Onsan-eup, Ulju-gun, Ulsan, 44992, Republic of Korea
[e] Department of Energy Engineering, School of Energy and Chemical Engineering, UNIST, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, Republic of Korea

ARTICLE INFO

Keywords:
Sensor
Machine learning
Chemical property
Feature
Classification performance

ABSTRACT

Chemical recognition using machine learning based on detection by gas sensors relies on the accuracy and sensitivity of the sensors at capturing the key features of target classes. In some cases, however, the electronic signal transduced from the detection of analytes does not completely represent the key attributes, resulting in inaccurate classification results when trained from signal data alone. To overcome this shortcoming, we propose a novel "chemistry-informed" machine learning framework composed of two modules. From available sensor response data, Module 1 identifies and predicts the chemical properties of the analytes that give rise to the sensitivity and selectivity of the sensors, and Module 2 performs final classifications using the dataset concatenating predicted chemical properties and raw sensor responses. To evaluate the performance and generalizability of our methodology, we conducted experiments with three gas sensor array datasets for gas detection. In all the cases, the performance of gas species classification was improved when the raw features were combined with the predicted chemical property features. The main contribution of our framework is that it bridges the gap between the gas sensor signals and the target analytes, thereby improving classification performance beyond that of models trained exclusively on sensor response data.

1. Introduction

A chemical sensor is a device that transforms chemical information, ranging from the concentration of a particular component to a complete compositional analysis, into an analytically useful signal [1]. Gaseous analytes need to be detected in numerous fields including medical diagnostics, food quality control, industrial monitoring for safety, household applications, and environmental monitoring [2]. Data-driven approaches to chemical sensing have become essential technologies for modern industries emerging in the fourth industrial revolution [3–5]. Numerous attempts have been made to apply machine learning based on signal data obtained from chemical sensors to extract useful patterns from the electrical signal to predict the information of analytes. For example, Jurs et al. reviewed supervised and unsupervised machine learning techniques for pattern recognition using chemical sensors [6]. Pardo et al. proposed a classification method using hybrid sensor array data based on feature selection by principal component analysis (PCA) [7]. Cho et al. applied deep learning to detect low-concentration analytes using chemical sensors [3]. Ye et al. used the alternating noise spectrum of gas sensors to enhance classification power for chemically and structurally similar gases [8]. Krivetskiy et al. improved the selectivity of gas detection using statistical shape analysis preprocessing instead of conventional signal preprocessing methods [9]. Acharyya et al. enhanced the selectivity of sensing by applying high sensitive microstructure, and using classification models in which the multiple features such as temperature, response, and concentration are fed in Ref. [10]. They also applied gas sensing kinetics, and used fitted parameters for discrimination of VOCs [11]. A competition for predicting the intensity and pleasantness of odors from different molecules was held recently [12].

Detecting the variations in sensor signal patterns generated by different analytes is critical for successful chemometric analysis, particularly for gas classification using sensor data. Among the various attributes of signals and techniques, the magnitude of the change in signal intensity is most commonly extracted and used for pattern

[*] Corresponding author. Department of Industrial Engineering, UNIST, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, Republic of Korea.
E-mail address: chlim@unist.ac.kr (C. Lim).

https://doi.org/10.1016/j.chemolab.2023.104808
Received 29 August 2022; Received in revised form 13 February 2023; Accepted 10 March 2023
Available online 17 March 2023

Abbreviations

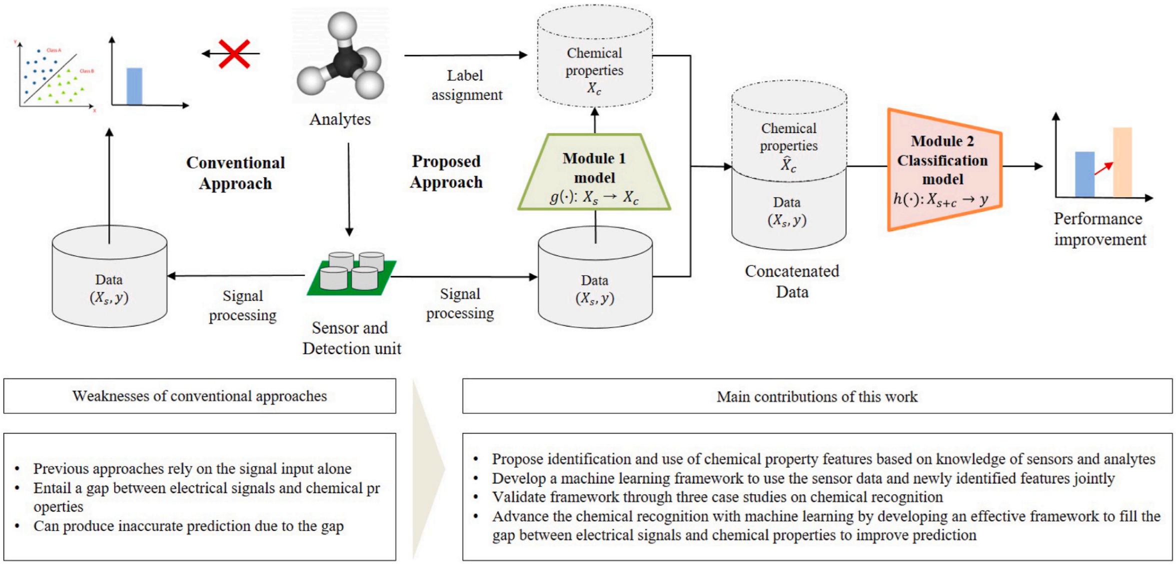
ADC	analog–digital converter
BA	butyl acetate
BD	mixture of butyl acetate and dimethyl sulfide
BDE	bond-dissociation energy
C	carbon monoxide
CE	mixture of carbon monoxide and ethylene
CNN	convolutional neural network
D	dimethyl sulfide
Ea	Ethanol
EDE	electron-donation effect
EM	mixture of ethylene and methane
Ey	ethylene
GBM	Gradient Boosting Machine
IP	Relative abundance of substances with low ionization potential
KNN	K-nearest neighbors
Me	methane
MOX	metal oxide
PCA	principal component analysis
PID	photoionization detector
SHAP	Shapley Additive exPlanations
SUL	the concentration of sulfur or sulfide in a sample
SVC	support vector classifier
SVM	support vector machine
SVR	support vector regressor
T	toluene
TB	mixture of toluene and butyl acetate
TD	mixture of toluene and dimethyl sulfide
TDB	mixture of toluene, dimethyl sulfide, and butyl acetate
TSEI	Taesung Environmental Research Institute
LightGBM	Light Gradient Boosting machine

recognition, which relies strongly on the concentration and unique physicochemical characteristics of the analyte [13]. In particular, specific chemical properties of analytes lead to different chemical reactions on the surfaces of gas sensors attuned to those analytes, generating characteristic output signals that form the foundation of gas sensor selectivity. Therefore, these signals contain information that is important for explaining the differences between analytes.

However, intrinsic gaps remain between a sensor's signal and an analyte's chemical properties because in a real-world application of gas classification, the sensor response does not directly represent the actual chemical properties of the analyte. Instead, it represents the combined output of the concentration, chemical properties of the analyte, and external factors. Because targets in machine learning models are defined on the basis of observable differences between substances having dissimilar chemical properties or concentrations, predictions made exclusively using sensor response data inherently involve the noise and information loss created by the gaps in the electrical sensing of chemicals. Consequently, this limited and naïve approach may provide inaccurate results.

In this study, we propose a novel chemistry-informed machine learning framework for gas classification that is designed to bridge the gap between observable gas sensor responses and unobservable chemical properties of target analyte classes inherent in the sensor response. As shown in Fig. 1, the framework consists of two modules: a regression or classification model that predicts chemical properties to identify new features and concatenates them (Module 1), and a final classifier with concatenated features of predicted chemical properties and raw sensor responses (Module 2). Module 1 requires knowledge-based identification of chemical property features which are chemically explainable and are linked to the sensor response. Based upon this information, the module predicts the presence (by classification) or magnitude (by regression) of chemical properties. Then, Module 2 performs classification for predicting target substances using the concatenated data that includes raw features and newly predicted features. In short, the framework incorporates domain knowledge to generate intermediate outputs (i.e., chemical knowledge about sensors and analytes, which can be estimated or calculated but cannot be directly measured by the sensors) that are used to improve the final classification performance from sensor data.

To evaluate the performance and generalizability of our method, we made empirical use of three gas datasets: a twin gas sensor arrays dataset [14] (Gas dataset 1), data from a gas sensor array exposed to turbulent gas mixtures [15] (Gas dataset 2), and an experimental gas sensor array dataset obtained from the Taesung Environmental Research Institute


Class A
Class B
Analytes
Label assignment
Chemical properties X_c
Conventional Approach
Proposed Approach
Module 1 model $g(\cdot): X_s \to X_c$
Chemical properties $\hat{X}_c$
Data (X_s, y)
Module 2 Classification model $h(\cdot): X_{s+c} \to y$
Performance improvement
Concatenated Data
Data (X_s, y)
Signal processing
Sensor and Detection unit
Signal processing
Data (X_s, y)
Weaknesses of conventional approaches
Main contributions of this work

- Previous approaches rely on the signal input alone
- Entail a gap between electrical signals and chemical pr operties
- Can produce inaccurate prediction due to the gap

- Propose identification and use of chemical property features based on knowledge of sensors and analytes
- Develop a machine learning framework to use the sensor data and newly identified features jointly
- Validate framework through three case studies on chemical recognition
- Advance the chemical recognition with machine learning by developing an effective framework to fill the gap between electrical signals and chemical properties to improve prediction

Fig. 1. Overview and contribution of this work.

(TSEI) in South Korea (Gas dataset 3). Considering that gas sensor outputs are strongly influenced by the concentrations and chemical properties of the analytes, these are suitable experimental datasets. However, although the aforementioned datasets contain sensor response data, they have relatively small populations. To overcome this, we extracted chemical property information from a small number of samples and utilized these properties as new features. We demonstrated the improvement in classification performance when chemical property features are concatenated with raw sensor features.

In relation to our work, two conventional methodologies—feature engineering and knowledge-based learning—have been used to minimize inaccurate results. However, those methods are incapable of retrieving important but hidden information buried beneath the original data. More rigorous feature engineering techniques have been emphasized as a way to improve the robustness of prediction models [16–18]. Numerous studies beyond the scope of chemical sensors have utilized feature engineering techniques to enhance the performance of prediction models. For example, Schroeder et al. performed feature selection by determining the optimal features via independent selector classification and a combinatorial scan of all possible feature combinations for food classification from sensor array data [19]. Diaz et al. proposed a feature selection technique, which selects generalizable variables across domains [20]. Varzaneh et al. proposed a feature selection technique using Entropy-based and Lévy flight to improve the performance of Equilibrium Optimization, thereby enhancing classification performance [21]. Acharyya et al. proposed the Discrete wavelet transform (DWT) based preprocessing techniques to achieve selective discrimination of VOCs [22].

Other than chemistry fields, some methodological studies have developed feature engineering methods for practical use with existing data to enhance prediction performance. For instance, Liu et al. [23] proposed a feature generation method for enhancing the accuracy of predicting click-through rates using a convolutional neural network (CNN), and Chen et al. proposed a method of integrating heterogeneous features for wind power prediction using a predictive stacked autoencoder [24]. Zhao et al. proposed a two-module framework, including multi-modal neural networks and sparse-group LASSO for grouped heterogeneous feature selection, designed to improve classification performance [25]. DeVries and Taylor proposed a data-augmentation method based on feature spaces instead of input spaces using an autoencoder [26]. Wang et al. attempted to augment features based on logarithm marginal density ratio transformations [27]. As demonstrated by these studies, the literature of feature engineering has focused mainly on feature extraction, selection, and embedding. However, interpreting the outcomes of feature extraction is difficult. Feature selection inherently involves loss of information, and the classification performance depends on the selection criteria [27]. Essentially, existing studies focus on maximally exploiting the information in the original raw features. However, our work attempts to extend the available information *per se*.

Meanwhile, attempts have also been made to import external scientific knowledge into machine learning models to improve the machine learning performance. For example, physics-informed neural networks have been proposed [28], which utilize existing domain knowledge represented by partial differential equations as regularization terms in loss functions. Such models aim to enhance prediction performance by approximating the physical laws governing the domains of the data. However, although this approach enhanced the regression performance, it required large datasets to train neural networks. In addition, this approach regularizes a loss function that utilizes governing laws of physics in the form of ordinary differential equations, which are unavailable or unrealistic to derive in many cases. A different approach was proposed [29] to predict the compressive strengths of alkali-activated materials. It used simple feature engineering to construct the derived variables that represented chemical properties. Another study used given physics and chemical property features to predict the properties of glass [30]. However, these existing approaches used calculation-based feature aggregation to prepare chemical property features used in the final prediction. However, in many cases, particularly for sensors, calculating the chemical properties from sensor responses is complicated or difficult in real-time inference problems.

In contrast, our chemistry-informed machine learning method attempts to predict and use hidden but significant chemical information that influences gas sensor selectivity, but has not been used to date, thereby expanding the list of features beyond those available from domain knowledge. The hidden properties identified and utilized by our model are not directly reported during the electrical sensing of chemicals, nor are they calculable from conventional statistical methods. Therefore, we identified and predicted these properties in Module 1, and used them as additional features (dimensions) in the final predictions. Because the framework entails domain-knowledge-based labeling of the new chemical property features before they are concatenated with the original datasets with sensor response features, we can alleviate the noise and information loss from the gaps in electrical sensing of chemicals and expand the available information for machine learning. To the best of our knowledge, this is the first study that explores and exploits chemical property features for gas classification. Specifically, we aimed to utilize both electrical signals and chemical property features to enhance classification performance. As shown in Fig. 1, our main contributions are as follows. Recognizing the importance of chemical recognition using machine learning, we developed a knowledge-based method for feature identification to improve sensor data classification performance. Although this study illustrates three application cases with small sized training datasets, we show that the proposed framework can improve the performance of gas recognition. The framework is applicable to the gas recognition system where the domain knowledge on training analytes are available, even when the number of gas species to recognize is limited to the certain scope, such as the e-nose implemented on food production facilities for detecting the spoilage of certain toxic chemical products.

2. Material and methods

In this section, we describe our framework for the identification of chemical property features, including Modules 1 and 2, in detail. Table 1

Table 1
Notation.

Name	Meaning
$X_s \in \mathbb{R}^{n\times m}$	Matrix including sensor response of all features
$C \in \mathbb{R}^{n\times q}$	Concentration matrix of mixture components (ppm)
$P \in \mathbb{R}^{l\times q}$	Chemical property matrix of mixture components (per unit component)
$X_c \in \mathbb{R}^{n\times l}$	Total chemical property matrix for mixture
$\widehat{X}_c \in \mathbb{R}^{n_{test}\times l}$	Features generated by Module 1
$X_{s+\hat{c}} \in \mathbb{R}^{n_{test}\times (m+l)}$	Concatenation of X_s and $\widehat{X}_c$ data
$X_{s+c} \in \mathbb{R}^{n\times (m+l)}$	Concatenation of X_s and X_c data
$X_{s+random} \in \mathbb{R}^{n\times (m+l)}$	Concatenation of X_s and l random Gaussian noise feature
$x_{s,i} \in \mathbb{R}^{n}$	A vector including response from sensor i for all samples ($i =$ 1 to m)
$x_{c,j} \in \mathbb{R}^{n}$	A vector including chemical property j for all samples ($j =$ 1 to l)
$y \in \mathbb{R}^{n}$	Target variable (Gas species)
L	Loss function: Mean squared error (L_{MSE}), Cross entropy loss (L_{CE})
l	Number of chemical property features
m	Number of sensors
n_{train}	Training set size
n_{test}	Test set size
n	Number of samples ($n = n_{train} + n_{test}$) ($k = 1$ to n)
q	Number of components in a mixture of analytes
θ	Model parameters
ε	Error vector
$f(\bullet)$	Sensor response function
$g(\bullet) : X_s \rightarrow X_c$	Module 1 function
$h(\bullet) : X_{s+c} \rightarrow y$	Module 2 function

details the notation used in this study, and Fig. 2 shows an overview of the architecture of the framework.

2.1. Module 1

The primary function of Module 1 is to extract inherent chemical property features from sensor response X_s, which is accomplished by pre-defining hidden chemical property labels X_c and obtaining predicted properties $\widehat{X}_c$. The identification of chemical properties involves tracing sensor signals back to ascertain their origin. For example, the selectivity and sensitivity of metal-oxide gas sensors are strongly influenced by the specific traits of each sensor in the array and the quantities and unique chemical properties of the analytes [31].

In short, Module 1 focuses on capturing the relationship between the given sensor response (X_s) and the hidden chemical properties of sensors and analytes (X_c). This can be accomplished using two typical supervised learning methods: regression and classification. The choice between regression (numerical labels) and classification (categorical labels) depends on the user's definition of X_c. In this study, we implement both techniques and encourage researchers and engineers to select the appropriate strategy for their purpose. We implemented regression for two gas sensor array datasets obtained from UCI machine learning repository [14,15] and classification for the TSEI gas sensor array dataset (see Section 3 for further details).

Following this prediction step (Module 1), the predicted value $\widehat{X}_c$ of the test set is used for the inference phase in Module 2 because X_c has pre-defined labels and is non-observable in the test-and-use phases.

Meanwhile, in terms of the choice of Module 1 model to predict X_c from X_s, for gas dataset 1 and 2 using continuous X_c variables, we used LightGBM regressor, because, in contrast to conventional GBM, LightGBM prevents overfitting via strong regularization while preserving its performance and provides fast computations [32]. For gas dataset 3 using discrete X_c variables, we used random forest classifier, as it is an effective machine learning technique for discretely distributed data. For continuous data, an efficient discretization algorithm is applied prior to the learning step [33,34].

2.1.1. Regression for Module 1

A regression-type model is eligible for Module 1 if X_c is defined to be in continuous space. Several studies have constructed quadratic models for quantifying and representing the sensor response from analyte concentrations in a mixture [35–37]. Therefore, we define the signal function to include concentration C and chemical properties P summed with error ε, which consists primarily of noise from external environmental factors, as

$$x_{s,i} = f_i\left(c_1, c_2, \dots, c_q, p_1, p_2, \dots, p_l, \right) + \varepsilon \tag{1}$$

where x_{si} is an element of X_s, which consists of the combination of feature vectors of sensor responses:

$$X_s = \left[x_{s,1}, x_{s,2}, \dots, x_{s,m}\right] \tag{2}$$

For simplicity, we represent the chemical property matrix X_c as a linear combination of concentration C and unit chemical properties P, with an error term. For the parameter-based model, X_c is denoted as in Eq. (3).

$$X_c = CP^T = g\left(X_s \middle| \theta_g\right) + \varepsilon \tag{3}$$

X_c and $\widehat{x}_{c,j}$ can also be denoted as

$$X_c = \left[x_{c,1}, x_{c,2}, \dots, x_{c,l}\right] \in \mathbb{R}^{n \times l} \tag{4}$$

$$L_{MSE,\ j} = \frac{1}{n_{train}} \sum_{k=1}^{n_{train}} \left(x_{c,k,j} - \widehat{x}_{c,k,j}\right)^2 \tag{5}$$

$$\widehat{x}_{c,j} = g_j\left(X_{s,test} \middle| \theta^*_{g,j}\right) \in \mathbb{R}^{n_{test}} \tag{6}$$

where $g_j(\bullet) : X_s \rightarrow x_{c,j}$ denotes the prediction model for jth chemical property in Module 1, whose independent and dependent variables are X_s and $x_{c,j}$, respectively, and $\theta^*_{g,j}$ denotes the optimal parameter set of $g_j(\bullet) : X_s \rightarrow x_{c,j}$. $\theta^*_{g,j}$ is obtained by minimizing the loss function, i.e., the mean-squared error $L_{MSE,j}$ (5).

2.1.2. Classification for Module 1

For classification applications, discrete one-hot labels $x_{c,k,j}$ are defined and assigned by users according to the presence or abundance of analytes having particular chemical properties (i.e., C and P). Therefore, the element of X_c is labeled as

$$x_{c,k,j} = \begin{cases} a_0 \text{ if} \\ \text{condition 0} \\ \vdots \\ a_r \text{ if} \\ \text{condition r} \end{cases} \tag{7}$$

where k and j range from 1 to n and l, respectively. Thereafter, for a parameter-based model, $\widehat{x}_{c,j}$ is obtained by minimizing the cross-entropy loss $L_{CE,j}$ (8):

$$L_{CE,j} = -\sum_{k=1}^{n_{train}} x_{c,k,j} \log\left(\widehat{x}_{c,k,j}\right) \tag{8}$$

$$\widehat{x}_{c,j} = g_j\left(X_{s,test} \middle| \theta^*_{g,j}\right) \in \mathbb{R}^{n_{test}} \tag{9}$$

2.2. Module 2

Module 2 is the final classifier for predicting the target variable y. For this process, $\widehat{X}_c$, the predicted chemical property of the test set, is obtained from Module 1, and concatenated with sensor response data X_s to

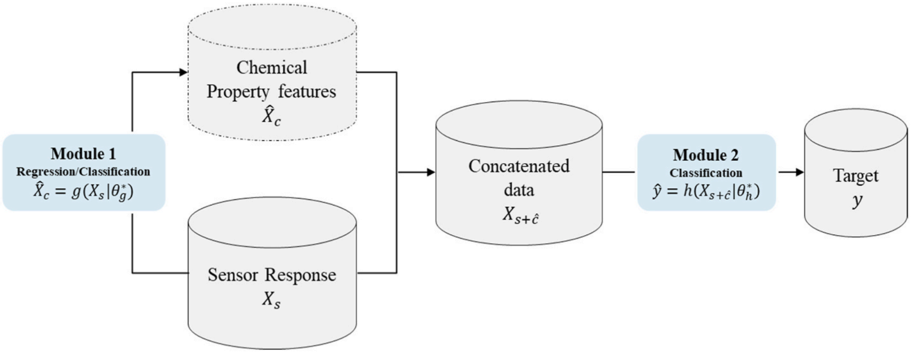


Fig. 2. The proposed framework for chemistry-informed machine learning.

obtain $X_{s+\hat{c}}$, which includes the sensor responses as well as the predicted chemical properties of the samples:

$$X_{s+\hat{c}} = \left[x_{s,1,test}, x_{s,2,test}, \ldots, x_{s,m,test}, \widehat{x}_{c,1,test}, \widehat{x}_{c,2,test}, \ldots, \widehat{x}_{c,l,test}\right] \in \mathbb{R}^{n_{test}\times(m+l)} \tag{10}$$

Module 2 acts on these concatenated data according to

$$\widehat{y} = h\left(X_{s+\hat{c}}\middle|\theta_h^*\right) \in \mathbb{R}^{n_{test}} \tag{11}$$

where $h(\bullet) : X_{s+c} \rightarrow y$, whose classification algorithms work similarly to those of Module 1.

Algorithm 1 presents the pseudocode summarizing the framework, including both modules.

Algorithm 1. Pseudocode for identification of chemical property features and classification

```
procedure (Module 1 and Module 2)
    Input: Training set ($X_{s,train}, X_{c,train}, y_{train}$); Test set ($X_{s,test}$); Module 1 $g(\bullet|\theta_g) : X_s \rightarrow X_c$; Module 2 $h(\bullet|\theta_h) : X_{s+c} \rightarrow y$;
    Output: Best model $h(\bullet|\theta_h^*)$
    for j = 1→l do
        Train $g_j(X_s|\theta_{g,j})$ with $X_{s,train}$ and $x_{c,j,train}$
        $\theta_{g,j}^* = \underset{\theta_{g,j}}{argmin} L(g_j(X_{s,train}), x_{c,j,train})$
        $\widehat{x}_{c,j} = g_j(X_{s,test}|\theta_{g,j}^*)$
    end for
    $X_{s+c,train} \leftarrow concatenate(X_{s,train}, X_{c,train})$
    $X_{s+\hat{c},test} \leftarrow concatenate(X_{s,test}, \widehat{X}_{c,test})$
    Train $h(X_{s+c,train}|\theta_h)$ with $X_{s+c,train}$ and $y_{train}$
    $\theta_h^* = \underset{\theta_h}{argmin} L_{CE}(h(X_{s+c,train}), y_{train})$
    Test $\widehat{y} = h(X_{s+\hat{c},test}|\theta_h^*)$
    return $\widehat{y}$
end procedure
```

2.3. Dataset descriptions

In this section, we describe the experimental procedures and validation results of our proposed framework with three datasets: 1) twin gas sensor arrays dataset [14], 2) data from a gas sensor array exposed to turbulent gas mixtures, which are publicly available in the UCI machine learning repository [15], and 3) a gas sensor array dataset obtained from TSEI (Ulsan, South Korea). Using the three gas sensor datasets, we finally predict the presence of gas components of unknown gas mixtures. The details of the datasets are summarized in Table S9.

2.3.1. Twin gas sensor arrays dataset

Gas dataset 1 [14] was collected from a sensor array consisting of 8 MOX sensors (Figaro USA, Inc., Glenview, IL, USA) of four types. The experimental protocol consisted of 5 chemical detection units having the same experimental setting, and each sensor unit was tested several times over a period of 22 days. For each test, four different target volatiles (Ethylene, Ethanol, Carbon Monoxide, Methane) mixed with dry air was passed through a sensing chamber at 10 different concentrations, while the total flow rate was maintained at 400 mL/min. The total duration of each experiment was 600 s. Air was circulated through the sensing chamber for 50 s, whereafter the target gas was circulated for 100 s. Finally, at 150 s, the gas was purged, and clean air was circulated through the chamber. The total number of samples was 640, and they were equally distributed between the four classes (i.e., 160 samples per class). Detailed information is available in Fonollosa et al., 2016 [14].

2.3.2. Gas sensor array exposed to turbulent gas mixtures

Gas dataset 2 [15,38] was collected from chemo-resistive gas sensors exposed to turbulent gas mixtures of carbon monoxide, methane, and ethylene in a wind tunnel. The sensor array consisted of eight commercial MOX sensors (Figaro USA, Inc., Glenview, IL, USA) of six types, whose selectivity and sensitivity to target gases varied. The complete time-series dataset consisted of 180 sample measurements, and the total duration for each sample was 300 s. At $t = 60$ s, the gases were released and persisted for 180 s. In the final 60 s, the gases were purged and the system recovered. Tables S3 and S4 (Supplementary Material) list the details of the sensor array and the distribution of samples according to gas concentrations, respectively. See Fonollosa et al., 2014 [15] for detailed information.

2.3.3. TSEI dataset

From TSEI, we obtained 104 time-series data samples including saturation and recovery phases for mixtures of two or more gases (collected April–June 2020). The sensor array consisted of seven sensors, including MOX sensors, electrochemical sensors, and a photoionization detector (PID), whose manufacturers and target gases were different. Toluene (T), butyl acetate (BA), and dimethyl sulfide (D) were selected as the target gases because the dataset aims to build e-nose system which is served in refrigerators, and three gases are the most likely to cause the odors due to food spoilage. Tables S6 and S7 list the details of the sensor array and the distribution of samples according to gas concentrations, respectively, and Fig. S12 describes a schematic diagram of gas sensing setup of the experiment of TSEI dataset.

For each data sample, the total duration was 420 s. For the first 140 s, no gas flowed, and from $t = 141$ s, the gas mixture was flowed turbulently until all the sensors in the array were saturated. The signal was converted to 0–4000, corresponding to 0–5 V, using an analog–digital converter (ADC). The experiment was conducted at room temperature (25 ± 2 °C) and ambient humidity (60 ± 5% RH).

2.4. Data preprocessing

In this study, we focus on the extraction of chemical properties from maximum steady-state change, for application in real-world conditions. Maximum steady-state change is the most popular feature used in chemo-sensory data processing. Previously, Nallon et al. [39] extracted parameters by fitting the saturation recovery curve to reflect physicochemical relationships such as adsorption and desorption rates. In order to prepare X_s, we extracted three parameters from the sensor curve. The first and second parameter are the maximum change in resistance ($\Delta R = R_{\max} - R_{\min}$) or voltage ($\Delta V = V_{max} - V_{min}$), which are the raw feature in the time-series data for a sample because we assumed that the chemical properties X_c inherent in the analytes are most clearly and maximally revealed in the saturated steady-state phase, during which the flow of electrons through the circuit is maximized. The third parameter is the coefficient of the curve fitting the saturation phase. According to Nallon et al. [39], the saturation phase can be fitted to a curve Eq. (14), and α_s, the coefficient of the exponential term in the equation of the curve, is closely related to the molar mass of the analyte (see Section 2.5 for detailed explanation). Fig. 3 shows the saturation and recovery curve of the sensor signal, and the preparation of X_s.

Thereafter, we deleted some abnormal samples having unusual sensor data. Among various types of data preprocessing methods used in chemical sensor array response processing [6,40], we performed robust scaling for the as-obtained X_s as described in Eq. (12):

$$x_{s,k,i}^{scaled} = \frac{x_{s,k,i} - Q_2\left(x_{s,i}\right)}{Q_3\left(x_{s,i}\right) - Q_1\left(x_{s,i}\right)} \tag{12}$$

in which $Q_1, Q_2,$ and Q_3 represent the 1st, 2nd and 3rd quartile, respectively.

2.5. X_c preparation and Module 1

We identify five chemical property features in total. Among them, for sensors with ZnO/Al_2O_3 MOX sensors (Table S1), we defined three major chemical property candidates for X_c with which the selectivity of the sensors is closely related: the molar mass function of the analytes [39], the bond dissociation energy (BDE) [41], and the electron-donation effect [42] (Table 3). These properties were also applied to gas dataset 2 (see Section 3.2). For the other properties, Ionization potential (IP) and the concentration of sulfur or sulfide in a sample (SUL) are defined as binary variables, which are applied in gas dataset 3. Here, we listed the description of five properties.

- Bond dissociation Energy (BDE): The strengths of the chemical bonds existing in substances during chemical reactions are critical parameters because the extent to which a reaction proceeds is closely related to the energy required to break the original bonds in the reactants [41]. The smaller the BDE of the gas molecule, the more readily the redox reaction with O_2^- (ads) occurs on the sensor surface and the greater is the number of electrons likely to be released. Considering the tendency that high BDE causes low sensitivity, and low BDE causes high sensitivity, we simply defined p_{BDE} as the reciprocal of the BDE for each component (13).
- Molar mass function ($M^{-\frac{1}{2}}$): The molar mass function can be derived and estimated from the rate of adsorption. It is driven by the Langmuir kinetic equation (15), in which $(1-\theta)$ is the fraction of sites not covered, s^* is a sticking coefficient term, P_r is the pressure, N_A is the Avogadro constant, T is temperature, R is the gas constant, M is the molar mass, and E_{ads} is the activation energy of adsorption. According to Nallon et al. [39], Eq. (15) can be fitted onto hypothesis curve Eq. (14), in which α_s, β_s, γ_s are the trainable parameters of the saturation curve (Saturation phase on Fig. 3). Because the surface area and all other variables are fixed, the rate of adsorption varies with $\frac{1}{\sqrt{M}}$ and is strongly correlated with α_s, the coefficient of the exponential term in Eq. (14) [39].

$$p_{BDE} = \left[\frac{1}{BDE_1}, \frac{1}{BDE_2}, \ldots, \frac{1}{BDE_q}\right] \in \mathbb{R}^{1\times q} \tag{13}$$

$$R_s(t) = \alpha_s\left(1 - e^{-\beta_s t}\right) + \gamma_s \tag{14}$$

$$rate\ of\ adsorption = \frac{P_r N_A}{\sqrt{2\pi MRT}}(1-\theta)s^* e^{-E_{ads}/RT} \tag{15}$$

- Electron-donation effect (EDE): The number of electrons released per target gas molecule was calculated. The larger the EDE, the greater the number of electrons released during chemical reactions, which induces a stronger sensor response [42]. With different oxygen species (O^-, O_2^-, O^{2-}), O_2^- is mostly available on the sensor surface at room temperature, due to low activation energy [43–46]. Here, our calculation is based on using O_2^-, as all the experiments of our datasets were conducted at room temperature.

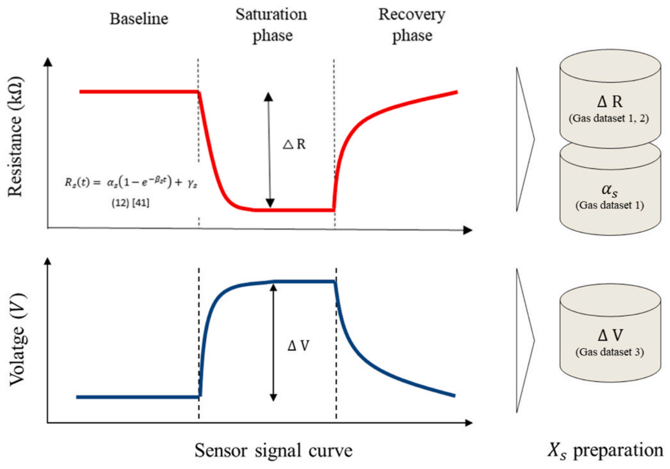


Fig. 3. The saturation and recovery curve of sensor signal and X_s preparation.

- Relative abundance of substances with low ionization potential (IP): The main advantage of the PID is selectivity toward gases having low ionization potentials [47], based on the difference between the ionization potential of the analytes (VOC) and photons ($h\nu$) released from the photon discharge source (a 10.6-eV krypton UV lamp), where h is Planck's constant ($h = 6.6262 \times 10^{-34} J \bullet s$), and ν is the frequency (Hz). If $h\nu$ exceeds the analyte's ionization potential, an excited molecule is produced (16), which subsequently produces a cation and an electron (17). The cations and electrons are drawn toward the cathode and anode respectively, generating current in the circuit [48–50].

$$VOC + h\nu \rightarrow VOC^* \tag{16}$$

$$VOC^* \rightarrow VOC^+ + e^- \tag{17}$$

In gas dataset 3, Sensor 3 is capable of detecting VOCs having ionization potentials lower than 10.6 eV (Table S6). The ionization potentials of its target gases: butyl acetate, dimethyl sulfide, and toluene are 10.01, 8.69, and 8.82 eV, respectively [53]. Although the ionization potential of butyl acetate (10.01 eV) is lower than the upper limit of detectability, the sensitivity was much lower than that for the other two target gases. Therefore, for kth sample ($k = 1$ to n), we assigned $x_{c,k,IP} = 1$ to samples whose total concentration of toluene and dimethyl sulfide exceeded 1000 ppm (76 of 104 samples) and assigned $x_{c,k,IP} = 0$ to the other samples.

- The concentration of sulfur or sulfide in a sample (SUL): A semiconductor sensor and an electrochemical sensor (e.g., Sensor 2 and 5 in gas dataset 3) capture sulfurous compounds. Since dimethyl sulfide is a sulfurous gas, $x_{c,k,SUF} = 1$ was assigned for the samples whose total concentration of dimethyl sulfide exceeded 500 ppm, and $x_{c,k,SUL} = 0$ to the remaining samples. For example, if a sample $x_{s,k}$ contains 1000, 2000, and 1000 ppm of dimethyl sulfide, toluene, and butyl acetate, respectively, the sum of concentration of dimethyl sulfide and toluene is 3000 ppm, and consequently, $x_{c,k,IP} = 1$, and $x_{c,k,SUL} = 1$ for the sample.

The process of preprocessing and X_s, X_c preparation, implementation specifications have been summarized in Table 2.

3. Result

In this section, we describe the experimental procedures and validation results of our proposed framework with three datasets: 1) twin gas sensor arrays dataset [14], 2) data from a gas sensor array exposed to turbulent gas mixtures, which are publicly available in the UCI machine learning repository [15,38], and 3) a gas sensor array dataset obtained from TSEI (Ulsan, South Korea). Using the three gas sensor datasets, we finally predict the presence of gas components of unknown gas mixtures. The details of the datasets are summarized in Table S9.

The random train/test splitting was performed as follows: the training set comprised 75% of the entire dataset, and the test set comprised the remaining 25% of it. We repeated the sampling for 100 times, shuffling the dataset each time. In the process, we stratified the classes, but did not consider the detection units. For both modules, we conducted Bayesian hyper parameter tuning (Table S10) for each repetition of random sampling via 3-fold cross-validation, then took the average of the test score for each task.

For Module 2, we focused on ascertaining the improvement created by our approach, comparing the performance of $h(X_s)$ and $h(X_{s+\hat{c}})$. We also compared the result of $h(X_{s+random})$, on which X_s and two or three random Gaussian noise features $x_{random,j} \sim N(0, 1^2)$ are concatenated

Table 2
Summary of preprocessing, X_s, X_c preparation, and Module 1.

Dataset	X_s	Total number of samples (before outlier deletions)	The number of deleted abnormal samples	Sensor types (The number of sensors in an array in parenthesis)	X_c (The total number of used features in parenthesis)	Module 1 model
Gas dataset 1	ΔR, α_s	640	1	MOX sensors (8)	BDE, EDE, $M^{-\frac{1}{2}}$ (3)	LightGBM regressor
Gas dataset 2	ΔR	180	4 [54]	MOX sensors (8)	BDE, EDE (2)	LightGBM regressor
Gas dataset 3	ΔV	104	–	MOX sensors (2), electrochemical sensors (4), PID sensors (1)	IP, SUL (2)	Random forest classifier

Table 3
Properties of target gases detected from MOX sensor: the property matrix was multiplied by the concentration matrix to obtain X_c.

Target gas	EDE (mol e^- released per mol gas) [42]	Molar mass function ($M^{-\frac{1}{2}}$) (mol/g)	BDE (kJ/mol) [51]	Surface redox reaction on sensing element [42] (At room temperature)
Carbon monoxide (CO)	0.5	0.189	1075	$2CO + O_2^-(ads) \rightarrow 2CO_2 + e^-$
Ethanol (C_2H_5OH)	2.5	0.147	436	$2C_2H_5OH \rightarrow 2CH_3CHO + 2H_2$ (basic oxide) $2CH_3CHO + 5O_2^-(ads) \rightarrow 4CO_2 + 4H_2O + 5e^-$ On the sensor surface, ethanol may undergo two different chemical reactions (dehydration and dehydrogenation) depending on the nature of the metal oxide. Because the sensors are composed of basic *ZnO*, dehydrogenation is favored [42,52]. Therefore, two ethanol molecules finally release five electrons after undergoing redox reactions on the surface of the sensor.
Ethylene (C_2H_4)	3	0.189	682	$C_2H_4 + 3O_2^-(ads) \rightarrow 2CO_2 + 2H_2O + 3e^-$
Methane (CH_4)	2	0.250	431	$CH_4 + 2O_2^-(ads) \rightarrow CO_2 + 2H_2O + 2e^-$

(Here, the number of $x_{random,j}$ is the same as the number of chemical property features: l). We report the F1 score as the evaluation metric because it is a sensitive tool for measuring classification performance in imbalanced datasets.

3.1. Twin gas sensor arrays dataset (Gas dataset 1)

In this experiment, we used three properties in total: the EDE and BDE obtained from input ΔR, and the molar mass function obtained from input α_s. The final total population of the training and test sets were 480 and 159, respectively. The distribution of samples is detailed in Table S2.

We predicted and generated $\widehat{X}_c$ for the test set, whereafter we calculated the mean and standard deviation of the correlation coefficient (R^2) of the test set. The results are presented in Table 4 and visualized in Fig. S4, which shows an example of the distribution of original X_c and generated $\widehat{X}_c$ for each model.

3.1.1. Module 2 and target prediction

For generating the classification models in Module 2, we used six different classifiers: elastic net, random forest, extra trees, LightGBM, K-nearest neighbors (KNN), and support vector classifier (SVC). We used ΔR instead of α_s as the input for Model 2 because ΔR and α_s extracted from the same sensor exhibited high correlation, which may have led to multicollinearity if both of them were provided as inputs to Module 2.

We performed a Bayesian hyperparameter search (see Table S11 for detailed information) and selected the hyperparameter set that produced the lowest loss, and finally tested it using the test set. As the evaluation metric, we used the F1 score, which is defined as the harmonic mean of precision and recall. Table 5 shows the average F1 score (and standard deviation thereof) of 100 random samplings for each model. We also present the win rate of F1 score, which is the number of win from 100 random train/test splittings. If multiple candidates gained the same highest score, then the win rate of all of them are counted.

As a result, the highest average F1 score of $h(X_{s+\hat{c}})$ was 0.910 (elastic net). Also, for five out of six models, $h(X_{s+\hat{c}})$ outperformed $h(X_s)$.

3.2. Data from gas sensor array exposed to turbulent gas mixtures (Gas dataset 2)

Using the LightGBM regressor, we predicted and generated $\widehat{X}_c$ for the test set and calculated the mean and standard deviation of Test R^2. The results are presented in Table 6 and visualized in Fig. S5, which shows an example of distribution of original X_c and generated $\widehat{X}_c$ for each model.

3.2.1. Module 2 and target prediction

Table 7 shows the average F1 score of 100 random samplings for each model and task. The highest average F1 score of three tasks were achieved for $h(X_{s+\hat{c}})$, with the values being 0.946 for carbon monoxide classification (SVC), 0.972 for ethylene classification (SVC), and 0.930 for methane classification (Random forest classifier).

3.3. TSEI gas sensor array dataset (Gas dataset 3)

We used random forest classifier to perform classification for Module 1. The final population of the training set was 76, and that of test set was 28, stratifying the components.

The F1 score was selected as the criterion for the binary classification of Module 1. Table 8 presents the results (mean and standard deviation) for the generation of $\widehat{X}_c$. The results indicate that the average F1 score of the level of ionization potential was 0.777, and that of sulfur or sulfide

Table 4
Module 1 results by model (standard deviations in parentheses).

Metric	EDE	BDE	Molar mass function ($M^{-\frac{1}{2}}$)
R^2	0.725 (0.044)	0.721 (0.048)	0.732 (0.047)

Table 5
F1 scores/win rate of Module 2 (standard deviations of F1 score in parentheses).

Input	Elastic net	Random forest	Extra trees	LightGBM	KNN	SVC
$X_{s+\hat{c}}$ (Ours)	**0.910/68 (0.022)**	0.900/27 (0.023)	**0.908/71 (0.020)**	**0.879/62 (0.029)**	**0.892/63 (0.025)**	**0.908/50 (0.023)**
X_s	0.895/28 (0.025)	**0.907/46 (0.024)**	0.892/16 (0.024)	0.869/26 (0.034)	0.882/39 (0.026)	0.907/50 (0.023)
$X_{s+random}$	0.890/11 (0.024)	0.904/37 (0.025)	0.894/24 (0.025)	0.869/25 (0.033)	0.794/0 (0.030)	0.890/5 (0.025)

was 0.847.

Upon completion of Module 1, the datatype of predicted $\hat{X}_c$ was converted to float, and it was delivered to Module 2.

3.3.1. Module 2 and target prediction

In Module 2, just as gas dataset 1 and 2, we compared the results of $h(X_s)$ and $h(X_{s+random})$ with those of $h(X_{s+\hat{c}})$, and the results are presented in Table 9, which show that average F1 score for $h(X_{s+\hat{c}})$ are superior to those of $h(X_s)$ and $h(X_{s+random})$, except for the classification of toluene using KNN. The highest F1 score of each task was obtained using $h(X_{s+\hat{c}})$, and the values were 0.791, 0.852, and 0.829 for the classification of butyl acetate (Extra trees classifier), dimethyl sulfide (Extra trees classifier), and toluene (LightGBM).

4. Discussion

4.1. Improvement in classification performance achieved with the proposed methodology

For the gas dataset 1, the highest average F1 score was achieved by elasticnet trained from concatenated $X_{s+\hat{c}}$. To interpret the model and determine the contributions of chemical property features, we implemented Shapley additive explanations (SHAP) [55,56], a summary plot whereof is presented in Fig. S9, which shows the average SHAP values of models obtained from 100 random samplings. The result indicates that the molar mass function ('1/sqrt(m)'), EDE and BDE were the first, third and fourth highest contributing features for the model, respectively, proving the dominant contribution of chemical property features.

To interpret the gas dataset 2, both BDE and EDE have been moderately contributed features for the model. For the classification of Carbon monoxide, which has the highest BDE, and the lowest EDE among three target species (Please see Table 3), both chemical property features ranked 3rd and 4th highest contributing ones among whole features. For the classification of the other gases, Ethylene and Methane have the highest EDE and the lowest BDE among three target gases, respectively. The results show that EDE for Ethylene and BDE for methane are ranked higher than the other one.

Table 6
Module 1 results by model (standard deviations in parentheses).

Metric	BDE	EDE
$\boldsymbol{R^2}$	0.616 (0.131)	0.528 (0.140)

Table 8
Module 1 results by model (standard deviations in parentheses).

Metric	The level of Ionization potential (IP)	The level of sulfur/sulfide (SUL)
F1 score	0.777 (0.101)	0.847 (0.072)

For gas dataset 3, we defined discretely distributed chemical property features: the ionization potential and the level of sulfur or sulfide. The former is designed to detect dimethyl sulfide and toluene, whereas the latter is designed to exclusively detect dimethyl sulfide. The results (Fig. S11) demonstrate that the SHAP impact of 'Sulfur/sulfide (scaled)' significantly contributed to all three tasks because it is capable of detecting the samples containing dimethyl sulfide at concentrations lower than 500 ppm. Additionally, 'Ionization potential (scaled)' contributed to the high classification performance for gases having low ionization potentials (i.e., dimethyl sulfide and toluene).

Our chemistry-informed machine learning framework performs better than conventional approaches because the chemical property features X_c are closely related to sensor signal and sensor selectivity, thereby bridging the gap between X_s and y (see Fig. 4). We identified the chemical property features X_c that are likely to be related with the target y, where each target class is composed of analytes having different chemical properties. Therefore, X_c facilitates the separation of classes and the training of classification models that can distinguish between adjacent data points generated by different classes with higher accuracy. In other words, for accurate and reliable chemical recognition, the variation of X_c complements that of X_s between classes. Consequently, it enhances the selectivity and sensitivity of chemical sensing and recognition.

4.2. Applicability of the proposed framework to other problems

We attempted to identify new features based on knowledge of the chemical properties of the sensors and analytes to establish the connections between sensor response data and the target class (Fig. 4). Our identification of new features was based on the following criteria to maximize the classification performance and prevent adding unnecessary features. Firstly, the new features should be known to be inherent in the sensor signal, leading to variations of signal response. Secondly, the variation of signal and chemical property features should explain the target class. To achieve this goal, we referred to numerous previous

Table 7
F1 scores/win rate of Module 2 (standard deviations in parentheses).

Input	Elastic net	Random forest	Extra trees	LightGBM	KNN	SVC
a. Classification of carbon monixoide						
$X_{s+\hat{c}}$ (Ours)	**0.945/88 (0.033)**	**0.920/72 (0.048)**	**0.931/82 (0.049)**	0.844/69 (0.054)	0.898/54 (0.048)	**0.946/81 (0.036)**
X_s	0.936/59 (0.035)	0.910/51 (0.049)	0.911/40 (0.047)	**0.855/82 (0.055)**	**0.905/60 (0.041)**	0.943/75 (0.038)
$X_{s+random}$	0.936/59 (0.035)	0.910/52 (0.050)	0.908/40 (0.063)	0.854/82 (0.055)	0.905/60 (0.042)	0.942/76 (0.038)
b. Classification of ethylene						
$X_{s+\hat{c}}$ (Ours)	**0.959/49 (0.039)**	**0.922/45 (0.048)**	**0.913/42 (0.055)**	**0.944/38 (0.051)**	0.935/40 (0.039)	**0.972/46 (0.032)**
X_s	0.956/47 (0.048)	0.917/41 (0.053)	0.909/40 (0.063)	0.937/33 (0.043)	0.936/41 (0.045)	0.971/46 (0.033)
$X_{s+random}$	0.956/47 (0.048)	0.918/41 (0.054)	0.908/40 (0.063)	0.936/34 (0.042)	**0.936/41 (0.045)**	0.971/46 (0.033)
c. Classification of methane						
$X_{s+\hat{c}}$ (Ours)	**0.918/89 (0.040)**	**0.930/86 (0.040)**	**0.903/90 (0.045)**	0.904/49 (0.042)	0.868/54 (0.044)	**0.921/78 (0.042)**
X_s	0.913/75 (0.042)	0.906/41 (0.044)	0.866/23 (0.055)	**0.921/65 (0.036)**	**0.877/70 (0.049)**	0.921/75 (0.040)
$X_{s+random}$	0.913/74 (0.039)	0.907/42 (0.045)	0.872/23 (0.057)	0.920/65 (0.037)	0.874/70 (0.059)	0.921/75 (0.040)

Table 9
Results of Module 2 (standard deviations in parentheses).

Input	Elastic net	Random forest	Extra trees	LightGBM	KNN	SVC
a. Classification of butyl acetate						
$X_{s+\hat{c}}$ (Ours)	**0.593/82 (0.115)**	**0.697/83 (0.142)**	**0.791/65 (0.103)**	**0.749/79 (0.124)**	**0.699/81 (0.113)**	**0.693/94 (0.120)**
X_s	0.458/29 (0.002)	0.658/57 (0.135)	0.754/45 (0.110)	0.675/37 (0.120)	0.576/17 (0.117)	0.457/8 (0.037)
$X_{s+random}$	0.456/26 (0.004)	0.612/36 (0.127)	0.491/2 (0.077)	0.654/30 (0.119)	0.531/18 (0.109)	0.457/5 (0.046)
b. Classification of dimethyl sulfide						
$X_{s+\hat{c}}$ (Ours)	**0.721/79 (0.139)**	**0.823/82 (0.128)**	**0.852/85 (0.113)**	**0.826/79 (0.128)**	**0.763/68 (0.118)**	**0.762/65 (0.100)**
X_s	0.570/20 (0.128)	0.785/48 (0.120)	0.696/13 (0.148)	0.711/15 (0.123)	0.728/42 (0.103)	0.692/43 (0.142)
$X_{s+random}$	0.621/33 (0.122)	0.769/41 (0.128)	0.690/13 (0.148)	0.706/19 (0.127)	0.708/30 (0.106)	0.720/56 (0.139)
c. Classification of toluene						
$X_{s+\hat{c}}$ (Ours)	**0.755/75 (0.091)**	**0.796/70 (0.113)**	**0.717/65 (0.156)**	**0.829/61 (0.102)**	0.727/29 (0.149)	0.694/40 (0.148)
X_s	0.696/38 (0.094)	0.782/64 (0.077)	0.543/22 (0.133)	0.828/59 (0.088)	**0.808/56 (0.085)**	0.628/24 (0.157)
$X_{s+random}$	0.704/40 (0.088)	0.739/40 (0.103)	0.592/39 (0.158)	0.817/47 (0.091)	0.757/26 (0.087)	**0.735/56 (0.14)**

studies, sensor manuals provided by manufacturers, and textbooks. For example, we calculated the number of electrons released during surface reactions, BDEs, and ionization potentials. In addition, we consulted domain experts from TSEI, and this work was based on extensive discussions with them. Furthermore, we obtained scanning electron micrographs and energy-dispersive X-ray spectra that provided compositional and morphological information for MOX sensors (Tables S1, S3, S6) to relate them to previous studies.

Considering the three examples based on the specific evidence suitable for each problem, we believe that the proposed framework can be applied similarly to other problems in different domains. As demonstrated in this study, new heterogeneous types of features closely related to the original features can be generated by user-defined supervised learning methods. Specifically, we improved the classification performance by 1) using both discrete and continuous chemical property features (using classification and regression, respectively), 2) comparatively evaluating the performance of various types of machine learning models, e.g., elastic net classifier, random forest, extra trees classifier, LightGBM, KNN, and SVC. We hope that our approach will facilitate new research into a wide range of prediction problems in other fields, such as, but not limited to, image detection [57], biosensors [58], electronic tongues for prediction of chemicals [59], and chromatographic analysis [60]. We also expect that users will be able to identify new features that are unused but may be exploitable by analyzing existing features and sensors using scientific knowledge. They can extract, select, and preprocess these according to their needs.

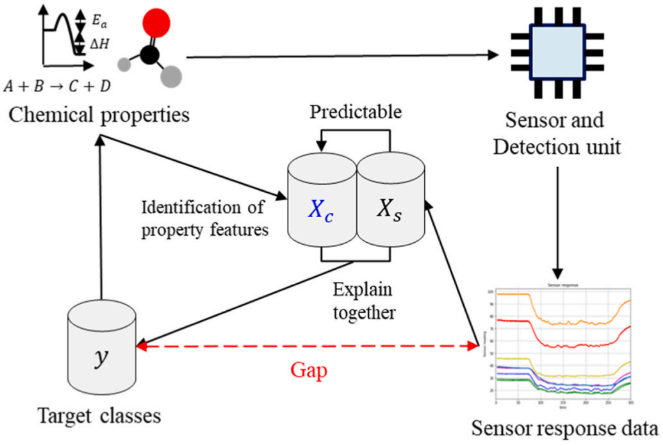


Fig. 4. Contribution of chemistry-informed machine learning: chemical property features bridge the gap between the target class and the sensor response data for effective prediction of analytes.

4.3. *Limitations and future work*

There are several limitations of this work that need to be addressed in future research. Firstly, in this study, we trained Modules 1 and 2 separately. However, simultaneous optimization of the two modules may be required, since the optimum result of Module 1 does not necessarily ensure the optimum test result in Module 2. Therefore, global optimization to improve both modules simultaneously is required. For example, in case we have large datasets, we can develop a chemistry-informed deep learning framework, where the Modules 1 and 2 are integrated into a neural network in an end-to-end manner (i.e., the error of Module 2 can be backpropagated to Module 1).

Secondly, our plan includes an attempt to develop a regression framework (Module 2) utilizing chemical property features. The scope of this study was confined to chemical recognition, and the key role of X_c in classification was to separate the classes. Different approaches, including feature analytics and extensive adoption of domain knowledge, may be required to enhance regression performance based on newly added chemical property features.

Finally, when time series sensor data are used for gas classification, we believe that a contrastive learning framework with chemical property features can be effective. There have been some recent studies to reflect and integrate domain knowledge onto a contrastive learning process (e.g. Ref. [61]). In the context of our work, gas sensing can be improved by using a knowledge-based objective function to learn time-series sensor array representation, as well as chemical property information.

5. Conclusion

Filling the knowledge gap between the data scientist and the chemist is a key step for ensuring success in applying machine learning models to real-world problems of chemical recognition. However, traditional approaches to machine learning for chemical recognition have relied on feature engineering of sensor response data instead of addressing the gap between sensor response and the sensed analytes. To improve machine learning performance with gas sensor data, we suggest identifying, predicting, and utilizing unused and unobserved but important chemical property features from observable sensor responses, based on domain knowledge. Here, domain knowledge refers to the chemical knowledge about sensors and analytes, which can be estimated or calculated but cannot be directly measured by the sensors. In particular, the “unique chemical properties of analytes” are affected by the selectivity of the gas sensors (i.e., using a specific sensor, some gas species are detected while others are not). In obtaining successful gas classification performance, the selectivity of gas sensors should be dealt importantly, as it makes the variants of the sensor signals for different gas species, and it strongly depends on the unique chemical properties of gas species, which “directly” affects the selectivity. Meanwhile, the estimation or calculation of such properties requires an understanding of the domain knowledge regarding the sensors and analytes, which is often not reflected in the sensor data *per se*.

Our bimodular "chemistry-informed machine learning" framework initially predicts chemical property features and subsequently performs classification using the concatenated features of predicted chemical properties and raw sensor responses. As such, our proposed framework is unique in that it helps the translation of the domain knowledge into specific features to improve the machine learning performance, which can be estimated using originally available features of sensor data. In improving the prediction performance using available data of the sensors, this domain-knowledge-based approach can be used to extend the features (from the set of original sensor features to the set of original features combined with chemical property features) and to enhance the performance of machine learning. As validated in this paper, the classification performance can be improved using this framework. Furthermore, once the modules 1 and 2 are well trained, these modules can be used to classify the gas in question automatically (i.e., after the training is completed, Module 1 can estimate the chemical property features with the original features, and Module 2 can classify the gas with the set of original features combined with chemical property features). Our framework can be generally applied to any gas datasets and any classification problems, as long as there is domain knowledge that can be translated into chemical property features.

Author statement

Yeram Kim: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization. Chiehyeon Lim: Conceptualization, Main Supervision. Junghye Lee: Supervision. Sungil Kim: Supervision. Sewon Kim: Data Curation. Dong-Hwa Seo: Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

This work was supported by the following institutes and grants: National Center for Materials Research Data (NCMRD) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. 2021M3A7C2089743), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A4A01049121), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (No. 2020-0-01336: Artificial Intelligence Graduate School Program - UNIST, No. 2021-0-02068, Artificial Intelligence Innovation Hub), and 2023 Project Fund (1.230058.01) of UNIST (Ulsan National Institute of Science & Technology).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2023.104808

References

[1] A. Hulanicki, S. Glab, F.o.l.k.e. Ingman, Chemical sensors: definitions and classification, Pure Appl. Chem. 63 (1991) 1247–1250, https://doi.org/10.1351/pac199163091247.

[2] G. Korotcenkov, Handbook of gas sensor materials, Conventional Approaches 1 (2013), https://doi.org/10.1007/978-1-4614-7165-3.

[3] S.Y. Cho, Y. Lee, S. Lee, H. Kang, J. Kim, J. Choi, J. Kim, Finding hidden signals in chemical sensors using deep learning, Anal. Chem. 92 (2020) 6529–6537, https://doi.org/10.1021/acs.analchem.0c00137.

[4] C. Lee, C. Lim, From technological development to social advance: a review of Industry 4.0 through machine learning, Technol. Forecast. Soc. Change 167 (2021), 120653, https://doi.org/10.1016/j.techfore.2021.120653.

[5] F.V. Paulovich, M.C.F. De Oliveira, O.N. Oliveira Jr., A future with ubiquitous sensing and intelligent systems, ACS Sens. 3 (2018) 1433–1438, https://doi.org/10.1021/acssensors.8b00276.

[6] P.C. Jurs, G.A. Bakken, H.E. McClelland, Computational methods for the analysis of chemical sensor array data from volatile analytes, Chem. Rev. 100 (2000) 2649–2678, https://doi.org/10.1021/cr9800964.

[7] M. Pardo, L.G. Kwong, G. Sberveglieri, K. Brubaker, J.F. Schneider, W.R. Penrose, J.R. Stetter, Data analysis for a hybrid sensor array, Sensor. Actuator. B Chem. 106 (2005) 136–143, https://doi.org/10.1016/j.snb.2004.05.045.

[8] H. Ye, C. Shi, J. Li, L. Tian, M. Zeng, H. Wang, Q. Li, New alternating current noise analytics enables high discrimination in gas sensing, Anal. Chem. 92 (2019) 824–829, https://doi.org/10.1021/acs.analchem.9b03312.

[9] V.V. Krivetskiy, M.D. Andreev, A.O. Efitorov, A.M. Gaskov, Statistical shape analysis pre-processing of temperature modulated metal oxide gas sensor response for machine learning improved selectivity of gases detection in real atmospheric conditions, Sensor. Actuator. B Chem. 329 (2021), 129187, https://doi.org/10.1016/j.snb.2020.129187.

[10] S. Acharyya, B. Jana, S. Nag, G. Saha, P.K. Guha, Single resistive sensor for selective detection of multiple VOCs employing SnO2 hollowspheres and machine learning algorithm: a proof of concept, Sensors Act. B: Chem. 321 (2020), 128484, https://doi.org/10.1016/j.snb.2020.128484.

[11] S. Acharyya, S. Nag, S. Kimbahune, A. Ghose, A. Pal, P.K. Guha, Selective discrimination of VOCs applying gas sensing kinetic analysis over a metal oxide-based chemiresistive gas sensor, ACS Sens. 6 (6) (2021) 2218–2224, https://doi.org/10.1021/acssensors.1c00115.

[12] A. Keller, R.C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, .P. Meyer, Predicting human olfactory perception from chemical features of odor molecules, Science 355 (2017) 820–826, https://doi.org/10.1126/science.aal2014.

[13] L.E. Kreno, K. Leong, O.K. Farha, M. Allendorf, R.P. Van Duyne, J.T. Hupp, Metal–organic framework materials as chemical sensors, Chem. Rev. 112 (2012) 1105–1125. https://doi.org/10.1021/cr200324t.

[14] J. Fonollosa, L. Fernandez, A. Gutiérrez-Gálvez, R. Huerta, S. Marco, Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization, Sensor. Actuator. B Chem. 236 (2016) 1044–1053. https://doi.org/10.1016/j.snb.2016.05.089.

[15] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, A. Vergara, R. Huerta, Chemical discrimination in turbulent gas mixtures with mox sensors validated by gas chromatography-mass spectrometry, Sensors 14 (2014) 19336–19353, https://doi.org/10.3390/s141019336.

[16] A. Zheng, A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly Media, Inc., 2018.

[17] J. Heaton, An empirical analysis of feature engineering for predictive modeling, in: SoutheastCon 2016 IEEE, 2016, pp. 1–6, https://doi.org/10.1109/SECON.2016.7506650.

[18] U. Khurana, D. Turaga, H. Samulowitz, S. Parthasrathy, Cognito: automated feature engineering for supervised learning, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE, 2016, pp. 1304–1307, https://doi.org/10.1109/ICDMW.2016.0190.

[19] V. Schroeder, E.D. Evans, Y.C.M. Wu, C.C.A. Voll, B.R. McDonald, S. Savagatrup, T. M. Swager, Chemiresistive sensor array and machine learning classification of food, ACS Sens. 4 (2019) 2101–2108, https://doi.org/10.1021/acssensors.9b00825.

[20] V.F. Diaz, P. Mishra, J.M. Roger, W. Saeys, Domain invariant covariate selection (Di-CovSel) for selecting generalized features across domains, Chemometr. Intell. Lab. Syst. 222 (2022), 104499, https://doi.org/10.1016/j.chemolab.2022.104499.

[21] Z.A. Varzaneh, S. Hossein, S. Ebrahimi, M.M. Javidi, A New Hybrid Feature Selection Based on Improved Equilibrium Optimization, Chemometrics and Intelligent Laboratory Systems, 2022, 104618, https://doi.org/10.1016/j.chemolab.2022.104618.

[22] S. Acharyya, S. Nag, P.K. Guha, Ultra-selective tin oxide-based chemiresistive gas sensor employing signal transform and machine learning techniques, Anal. Chim. Acta 1217 (2020), 339996, https://doi.org/10.1016/j.aca.2022.339996.

[23] B. Liu, R. Tang, Y. Chen, J. Yu, H. Guo, Y. Zhang, Feature generation by convolutional neural network for click-through rate prediction, in: The World Wide Web Conference, 2019, pp. 1119–1129, https://doi.org/10.1145/3308558.3313497.

[24] J. Chen, Q. Zhu, H. Li, L. Zhu, D. Shi, Y. Li, .Y. Liu, Learning heterogeneous features jointly: a deep end-to-end framework for multi-step short-term wind power prediction, IEEE Trans. Sustain. Energy 11 (2019) 1761–1772, https://doi.org/10.1109/TSTE.2019.2940590.

[25] L. Zhao, Q. Hu, W. Wang, Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso, IEEE Trans. Multimed. 17 (2015) 1936–1948, https://doi.org/10.1109/TMM.2015.2477058.

[26] T. DeVries, G.W. Taylor, Dataset Augmentation in Feature Space, 2017 arXiv preprint arXiv:1702.05538.

[27] H. Wang, J. Gu, S. Wang, An effective intrusion detection framework based on SVM with feature augmentation, Knowl. Base Syst. 136 (2017) 130–139, https://doi.org/10.1016/j.knosys.2017.09.014.

[28] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nat. Rev. Phys. 3 (2021) 422–440, https://doi.org/10.1038/s42254-021-00314-5.

[29] L.V. Zhang, A. Marani, M.L. Nehdi, Chemistry-informed machine learning prediction of compressive strength for alkali-activated materials, Construct. Build. Mater. 316 (2022), 126103, https://doi.org/10.1016/j.conbuildmat.2021.126103.

[30] Y.T. Shih, Y. Shi, L. Huang, Predicting glass properties by using physics-and chemistry-informed machine learning models, J. Non-Cryst. Solids 584 (2022), 121511, https://doi.org/10.1016/j.jnoncrysol.2022.121511.

[31] V.E. Bochenkov, G.B. Sergeev, Sensitivity, selectivity, and stability of gas-sensitive metal-oxide nanostructures, Metal Oxide Nanostruct. Their Applic. 3 (2010) 31–52.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, T.Y. Liu, Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. (2017) 30.

[33] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (3) (2002) 18–22.

[34] Y. Sun, H. Wang, B. Xue, Y. Jin, G.G. Yen, M. Zhang M, Surrogate-assisted evolutionary deep learning using an end-to-end random forest-based performance predictor, IEEE Trans. Evol. Comput. 24 (2) (2019) 350–364. https://doi.org/10.1109/TEVC.2019.2924461.

[35] P.K. Clifford, D.T. Tuma, Characteristics of semiconductor gas sensors I. Steady state gas response, Sensor. Actuator. 3 (1982) 233–254, https://doi.org/10.1016/0250-6874(82)80026-7.

[36] E. Llobet, X. Vilanova, J. Brezmes, J.E. Sueiras, R. Alcubilla, X. Correig, Steady-state and transient behavior of thick-film tin oxide sensors in the presence of gas mixtures, J. Electrochem. Soc. 145 (1998) 1772, https://doi.org/10.1149/1.1838556.

[37] S. Hirobayashi, M.A. Kadir, T. Yoshizawa, T. Yamabuchi, Verification of a logarithmic model for estimation of gas concentrations in a mixture for a tin oxide gas sensor response, Sensor. Actuator. B Chem. 92 (2003) 269–278, https://doi.org/10.1016/S0925-4005(03)00311-3.

[38] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, R. Huerta, Data set from chemical sensor array exposed to turbulent gas mixtures, Data Brief 3 (2015) 216–220, https://doi.org/10.1016/j.dib.2015.02.022.

[39] E.C. Nallon, V.P. Schnee, C.J. Bright, M.P. Polcha, Q. Li, Discrimination enhancement with transient feature analysis of a graphene chemical sensor, Anal. Chem. 88 (2016) 1401–1406, https://doi.org/10.1021/acs.analchem.5b04050.

[40] S.K. Jha, R.D.S. Yadava, K. Hayashi, N. Patel, Recognition and sensing of organic compounds using analytical methods, chemical sensors, and pattern recognition approaches, Chemometr. Intell. Lab. Syst. 185 (2019) 18–31, https://doi.org/10.1016/j.chemolab.2018.12.008.

[41] A. Mirzaei, H.W. Kim, S.S. Kim, G. Neri, Nanostructured semiconducting metal oxide gas sensors for acetaldehyde detection, Chemosensors 7 (2019) 56, https://doi.org/10.3390/chemosensors7040056.

[42] P. Rai, Y.T. Yu, Citrate-assisted hydrothermal synthesis of single crystalline ZnO nanoparticles for gas sensor application, Sensor. Actuator. B Chem. 173 (2012) 58–65, https://doi.org/10.1016/j.snb.2012.05.068.

[43] S. Acharyya, B. Manna, S. Nag, P.K. Guha, WO3 nanoplates based chemiresistive sensor device for selective detection of 2-propanol, IEEE SENSORS (2019) 1–4, https://doi.org/10.1109/SENSORS43011.2019.8956578.

[44] B.J. Wang, S.Y. Ma, S.T. Pei, X.L. Xu, P.F. Cao, J.L. Zhang, T. Han, High specific surface area SnO2 prepared by calcining Sn-MOFs and their formaldehyde-sensing characteristics, Sensors Act. B: Chem. 321 (2020), 128560, https://doi.org/10.1016/j.snb.2020.128560.

[45] S. Acharyya, S. Dey, S. Nag, P.K. Guha, ZnO cladded MnO 2 based resistive sensor device for formaldehyde sensing, IEEE SENSORS (2018) 1–4, https://doi.org/10.1109/ICSENS.2018.8589683.

[46] P. Bhat, S.K. Naveen Kumar, Evaluation of IDE-based flexible thin film ZnO sensor for VOC sensing in a custom designed gas chamber at room temperature, J. Mater. Sci. Mater. Electron. 33 (3) (2022) 1529–1541, https://doi.org/10.1007/s10854-021-07664-x.

[47] D.C. Locke, C.E. Meloan, Study of the photoionization detector for gas chromatography, Anal. Chem. 37 (1965) 389–395, https://doi.org/10.1021/ac60222a023.

[48] J.G. Sevcik, Detectors in Gas Chromatography, Elsevier, 2011.

[49] S.O. Agbroko, J. Covington, A novel, low-cost, portable PID sensor for the detection of volatile organic compounds, Sensor. Actuator. B Chem. 275 (2018) 10–15, https://doi.org/10.1016/j.snb.2018.07.173.

[50] Alphasense Ltd, PID-AH2 technical specification. https://www.alphasense.com/WEB1213/wp-content/uploads/2019/08/PID-AH2.pdf. (Accessed 18 April 2022).

[51] B. Ruscic, Active thermochemical tables: sequential bond dissociation enthalpies of methane, ethane, and methanol and the related thermochemistry, J. Phys. Chem. 119 (2015) 7810–7837, https://doi.org/10.1021/acs.jpca.5b01346.

[52] J. Xu, J. Han, Y. Zhang, Y.A. Sun, B. Xie, Studies on alcohol sensing mechanism of ZnO based gas sensors, Sensor. Actuator. B Chem. 132 (2008) 334–339, https://doi.org/10.1016/j.snb.2008.01.062.

[53] GfG. Instrumentation, Chemical ionization potential (eV) and 10.6eV PID correction factors (CF). http://goodforgas.com/wp-content/uploads/2013/12/TN2004_PID_gas_table_01_16_09.pdf. (Accessed 18 April 2021).

[54] L. Han, C. Yu, K. Xiao, X. Zhao, A new method of mixed gas identification based on a convolutional neural network for time series classification, Sensors 19 (2019) 1960, https://doi.org/10.3390/s19091960.

[55] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[56] L.S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. USA 39 (1953) 1095–1100, https://doi.org/10.1073/pnas.39.10.1095.

[57] M.E. Solmaz, A.Y. Mutlu, G. Alankus, V. Kılıç, A. Bayram, N. Horzum, Quantifying colorimetric tests using a smartphone app based on machine learning classifiers, Sensor. Actuator. B Chem. 255 (2018) 1967–1973, https://doi.org/10.1016/j.snb.2017.08.220.

[58] K.J. Squire, Y. Zhao, A. Tan, K. Sivashanmugan, J.A. Kraai, G.L. Rorrer, A.X. Wang, Photonic crystal-enhanced fluorescence imaging immunoassay for cardiovascular disease biomarker screening with machine learning analysis, Sensor. Actuator. B Chem. 290 (2019) 118–124, https://doi.org/10.1016/j.snb.2019.03.102.

[59] W.A. Christinelli, F.M. Shimizu, M.H. Facure, R. Cerri, O.N. Oliveira Jr., D. S. Correa, L.H. Mattoso, Two-dimensional MoS2-based impedimetric electronic tongue for the discrimination of endocrine disrupting chemicals using machine learning, Sensor. Actuator. B Chem. 336 (2021), 129696, https://doi.org/10.1016/j.snb.2021.129696.

[60] T. Vrzal, M. Malečková, J. Olšovská, DeepReI: deep learning-based gas chromatographic retention index predictor, Anal. Chim. Acta 1147 (2021) 64–71, https://doi.org/10.1016/j.aca.2020.12.043.

[61] M. T. Nonnenmacher, L. Oldenburg, I. Steinwart, D. Reeb, Utilizing expert features for contrastive learning of time-series representations. In International Conference on Machine Learning(pp. 16969-16989). PMLR.