chemosensors

Article

Unveiling Hidden Insights in Gas Chromatography Data Analysis with Generative Adversarial Networks

Namkyung Yoon †, Wooyong Jung † and Hwangnam Kim *

School of Electrical Engineering, Korea University, Seoul 02841, Republic of Korea; nkyoon93@korea.ac.kr (N.Y.); jy17347@korea.ac.kr (W.J.)
* Correspondence: hnkim@korea.ac.kr; Tel.: +82-2-3290-4821
† These authors contributed equally to this work.

Abstract: The gas chromatography analysis method for chemical substances enables accurate analysis to precisely distinguish the components of a mixture. This paper presents a technique for augmenting time-series data of chemicals measured by gas chromatography instruments with artificial intelligence techniques such as generative adversarial networks (GAN). We propose a novel GAN algorithm called GCGAN for gas chromatography data, a unified model of autoencoder (AE) and GAN for effective time-series data learning with an attention mechanism. The proposed GCGAN utilizes AE to learn a limited number of data more effectively. We also build a layer of high-performance generative adversarial neural networks based on the analysis of the features of data measured by gas chromatography instruments. Then, based on the proposed learning, we synthesize the features embedded in the gas chromatography data into a feature distribution that extracts the temporal variability. GCGAN synthesizes the features embedded in the gas chromatography data into a feature distribution that extracts the temporal variability of the data over time. We have fully implemented the proposed GCGAN and experimentally verified that the data augmented by the GCGAN have the characteristic properties of the original gas chromatography data. The augmented data demonstrate high quality with the Pearson correlation coefficient, Spearman correlation coefficient, and cosine similarity all exceeding 0.9, significantly enhancing the performance of AI classification models by 40%. This research can be effectively applied to various small dataset domains other than gas chromatography data, where data samples are limited and difficult to obtain.

Keywords: gas chromatography; generation model; deep learning

check for updates

Citation: Yoon, N.; Jung, W.; Kim, H. Unveiling Hidden Insights in Gas Chromatography Data Analysis with Generative Adversarial Networks. *Chemosensors* **2024**, *12*, 131. https://doi.org/10.3390/chemosensors12070131

Received: 29 May 2024
Revised: 2 July 2024
Accepted: 4 July 2024
Published: 7 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Chemical weapons pose a significant threat to global security and have been used in numerous conflicts around the world [1]. Chemical weapons are composed of microbiological and biological toxins consist of various types, such as nerve agents and blisters [2]. Their use has resulted in devastating consequences, including death, injury, and long-term health effects. The development of effective countermeasures and protective measures against chemical weapons is a critical area of research that requires an accurate and efficient analysis of chemical data [3]. Traditional chemical research methods, such as laboratory experiments and manual analysis, have limitations in terms of efficiency, accuracy, and scalability, especially when dealing with large datasets.

Among the various chemical analysis techniques to solve this problem, gas chromatography analysis methods are of particular importance for their ability to separate and accurately detect compounds within complex mixtures and are therefore indispensable for observing subtle changes in chemical properties, crucial for identifying and mitigating the threats posed by chemical weapons. This technique ensures not only accuracy but also detailed differentiation necessary for detecting harmful substances [4].

Artificial intelligence is being developed and studied in various fields such as detection, identification, and optimization [5–7]. In particular, recent advances in artificial intelligence

and machine learning offer significant potential to improve chemical research by enabling augmenting data for hard-to-test chemical experiments [8,9]. Such artificial intelligence has been widely applied to chemical analysis fields such as gas chromatography, showing promising results in improving research [10]. However, chemical data are very complex and structured, and limited quantities of experimental data are sometimes difficult to provide sufficient quantities of data for AI models, and AI models may have difficulty capturing the complex distributions of chemical properties [11].

Generative adversarial networks (GAN) are representative artificial intelligence techniques presented for the problems of limited data as described above [12]. A GAN is a type of neural network that consists of two parts: a generator and a discriminator. The generator generates synthetic data that is designed to look like real data, while the discriminator tries to distinguish between real and synthetic data. Through an iterative process, the generator learns to produce synthetic data that are increasingly difficult for the discriminator to distinguish from real data.

Training a GAN neural network requires a lot of training data. Data generation techniques using GAN are currently dominated by research for 2D data generation such as images and text generation models such as ChatGPT, which are currently in the spotlight, and also learn using giant corpus data [13]. In this paper, we propose and implement gas chromatography GAN (GCGAN) with specifically designed for chemical data to augment real chemical data. By enhancing chemical data through GCGAN, we aim to improve the accuracy and efficiency of chemical data analysis research beyond contributing to the development of effective countermeasures against chemical weapons.

Contributions of this paper: (1) Novel attention mechanism. Recently, with the development of transformer models such as ChatGPT, many attention mechanisms are being studied [14]. We propose a novel attention mechanism to adequately improve the performance of deep learning algorithms on gas chromatography data. Because of the characteristics of gas chromatography data, various existing deep learning techniques have enormous limitations in training. However, the *truncated* attention mechanism adequately learns about large peaks at retention time and small peaks at the rest of the time zone, which are properties represented by gas chromatography data. The discovery of the *truncated* attention mechanism suggests that it can be widely applied to a variety of data with similar properties, in addition to chemical data such as gas chromatography.

Contributions of this paper: (2) High-Performance GAN Architecture. In this paper, we design and fully implement a high-performance GAN architecture that generates chemicals using the mechanism of *truncated* attention. Currently, GAN has been studied to augment visual objects such as images. As many GAN studies focus on performance improvement, there are many learning time or limitations of learning, such as computation [15]. Therefore, our proposed GCGAN aims at efficient and high performance to address the above limitations. GCGAN represents a fusion structure using an autoencoder (AE) in addition to the attention mechanism of *truncated*. Through this, several advantages suitable for training chemical substances can be obtained. First, the training process of the generator is efficiently induced by reusing the AE model that learned useful features of the data. In addition, the transfer learning of the discriminator using AE improves the ability of the discriminator, allowing for more accurate classification of original and synthetic data. As such, GCGAN proposes the structure of a generative model that can not only generate high-quality simulation data of chemical data but also successfully learn the data in an efficient way.

Overall, our research is motivated by the need to overcome the limitations of traditional chemical research methods and to leverage the potential of AI and ML for improving chemical research. Our proposed novel attention mechanism and GAN structure offer a promising approach for augmenting chemical data and advancing chemical research.

The remainder of this paper is organized as follows. In Section 2, we review the existing literature on AI and ML in chemical research and current attention mechanisms and GAN structures. In Section 3, we propose new attention mechanisms and GAN

structures for chemical data. In Section 4, we present the results of our performance evaluation experiments using actual chemical data. Finally, in Section 5, we discuss the implications of our research and potential directions for future research.

2. Preliminary

2.1. Gas Chromatography Analysis

Gas chromatography is a widely used analytical technique in chemistry, especially in the field of organic chemistry. Gas chromatography separates and analyzes the components of a mixture based on their physical and chemical properties. In gas chromatography, the sample is vaporized and injected into a chromatographic column, where it interacts with a stationary phase. The components of the mixture separate on the basis of their interactions with the stationary phase and emerge from the column at different times, which are recorded as peaks on a chromatogram.

The resulting data from gas chromatography is a time series of signals, with each signal representing the concentration of a particular compound over time. The data are typically noisy, with fluctuations in the signal due to variations in the experimental conditions, such as temperature and flow rate. It is also easy to accurately identify and quantify individual compounds by peak [16]. Additionally, the data can be highly dimensional, with hundreds or thousands of signals collected for a single sample. In gas chromatography, the retention time refers to the time it takes for a compound to travel through the chromatographic column and elute from the detector. Retention time is an important parameter for identifying compounds, as it is influenced by the properties of the compound and chromatographic conditions. The retention time of a compound can be used as a fingerprint for identification, with different compounds exhibiting characteristic retention times. However, an accurate determination of the retention time can be challenging due to the presence of noise and other interfering compounds in the gas chromatography data.

Increasing data with GAN can improve the performance of artificial intelligence models in several ways [17]. In this work, we generate synthetic data to increase the size and diversity of gas chromatography datasets by generating synthetic data to improve the performance of artificial intelligence models.

In addition, the generated gas chromatography data can be used to filter noise or artifacts from the original gas chromatography data, resulting in cleaner and more reliable data for artificial intelligence models. The gas chromatography data generated in this way can help artificial intelligence models learn more meaningful features that are more useful for classification or prediction.

Additionally, the use of artificial intelligence, particularly attention mechanisms, can help accurately identify and quantify compounds based on retention time. Attention mechanisms can be applied to gas chromatography data to highlight the specific signals corresponding to the retention times of interest. This can aid in the identification and quantification of compounds, particularly in cases where the peaks are overlapped or the data are noisy. When the relevant signals are focused on, attention mechanisms can help to improve the accuracy and efficiency of gas chromatography data analysis.

Overall, the characteristics of gas chromatography data present significant challenges for traditional analytical techniques, such as identification. Therefore, the use of artificial intelligence, specifically generative adversarial neural networks and the attention mechanism, offers the potential to enhance the accuracy and efficiency of the analyzing of gas chromatography data.

The integration of artificial intelligence methodologies in the area of gas chromatography analysis is increasingly prominent, as evidenced by the spectrum of recent studies highlighted in comparison, as shown in Table 1. These studies predominantly utilize machine learning techniques, such as convolutional neural networks (CNN) and long-short-term memory networks, for tasks including peak classification, species authentication, and prediction of chromatography retention indices [9,18–21]. Each of these efforts has contributed significantly to enhancing the precision and efficiency of chemical analysis.

Table 1. Comparison of artificial intelligence studies on chromatography data.

Study	Methodology	Objective	Dataset	Outcome
Vrzal et al. [22]	Residual 2D-CNN model	Predict gas chromatography retention indices	Simplified molecular input line entry system (SMILES) data	Median percentage error $\leq$0.81%
AB Risum et al. [9]	CNN-based classification model	Evaluate GC-MS peaks	70,000 elution profile samples	AUC of 0.95 for peak classification
Y Wang et al. [18]	Various machine learning techniques	Authenticate species for improved safety and effectiveness	HPLC and GC analysis dataset	Achieved over 85% correct identification
DD Matyushin et al. [19]	Deep learning on polar and mid-polar phases	Predict gas chromatography retention indices	Polar phase dataset	Mean absolute error range of 16–50; outperforms existing methods
Tomas et al. [20]	LSTM-autoencoder	Predict liquid chromatography times	Dataset for molecular vectorization	Solvent labels predicted with 95.0% accuracy
Fedorova et al. [21]	1D-CNN model	Predict retention times in liquid chromatography	METLIN standard compounds and SMILES data	Mean absolute error of 34.7 s and median absolute error of 18.7 s
GCGAN (Our Work)	**GAN with a novel attention mechanism and transfer learning**	**Generate high-quality synthetic gas chromatography data**	**Own GC analysis dataset**	**Improve performance of AI model by more than 40%**

Moreover, the transformer structure, which is the basis for the LLM generation models represented by ChatGPT, is characterized by being suitable for Zipfian distribution and burstiness data. This is not suitable for chemical data analysis properties where values are uniformly measured over time. This characteristic underscores the suitability of the GAN-based model proposed in this paper for GC data augmentation, as supported by our previous research [11,23].

Distinguished from existing approaches, our work introduces GCGAN, leveraging a generative adversarial network enhanced by an innovative attention mechanism and transfer learning to conditionally generate high-quality synthetic GC data. This not only expands the available dataset, particularly beneficial where real samples are scarce, but also pioneers a novel avenue in the AI-chromatography domain by focusing on data generation.

2.2. GAN

GAN is a type of deep learning architecture that has gained significant attention due to their ability to generate highly realistic data samples. GAN consists of two neural networks: a generator (G) and a discriminator (D). The generator creates new data samples based on a random noise vector, while the discriminator evaluates the generated samples to determine whether they are real or fake.

Goodfellow et al. defines GAN as Equation (1), which is described as minimax game with value function $V(G, D)$ [12].

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \quad (1)$$

One of the key characteristics of GAN is their ability to learn from and mimic the underlying distribution of the training data. This is achieved through the iterative training process, where the generator learns to create increasingly realistic samples, while the discriminator becomes more skilled at identifying fake samples. As a result, GAN are capable of generating high-quality synthetic data that are often indistinguishable from the real data.

GAN have been widely used for generating images and text, but their potential applications in chemistry are only beginning to be explored [24]. In particular, GAN can be used to augment chemical data, which is important for tasks such as predicting chemical properties, designing new molecules, and identifying potential drug candidates.

One of the main advantages of using GAN for chemical data augmentation is their flexibility and adaptability [25]. GAN can be trained on a wide variety of data types, including images, audio, and text, which makes them well suited for use in chemistry, where data come in various forms. Moreover, GAN can be modified and customized to suit specific data types and applications. For instance, attention mechanisms can be incorporated into GAN architectures to improve their performance in analyzing and generating chemical data.

Another important characteristic of GAN is their ability to generate large quantities of data samples. This is particularly useful in chemistry, where data are often scarce and expensive to obtain. By augmenting the available data, GAN can improve the accuracy and reliability of machine learning models trained on chemical data.

In conclusion, we show the potential to contribute a lot to the development of the chemical field by enabling the generation of high-quality synthetic data through this paper. Furthermore, the ability to generate new chemical properties using GAN can greatly accelerate the drug discovery process and lead to the development of new and more effective drugs. However, further research is needed to develop new attention mechanisms and GAN architectures specifically tailored to chemical data. We propose the design of the appropriate artificial intelligence model through the subsequent Section 3 and demonstrate its performance through Section 4.

2.3. Attention Mechanism

The attention mechanism is a computational model that mimics the human cognitive system by selectively focusing on certain features or regions of input data, while filtering out irrelevant information [26]. In the context of chemical data, the attention mechanism is a type of neural network architecture that allows for the selective weighting of input features, based on their relevance to the task at hand [27]. Previously studied attention mechanisms are calculated using an attention score expressing the importance between input elements and an attention weight expressing how much each input element pays attention to another element. At this time, the attention score is performed by a method such as dot-product or additive, and indicates the importance or relevance of each element [28]. The attention score is then usually normalized using a softmax function to obtain up to 1 attention weight, by which the attention weight determines how much each element of the input sequence contributes to the output representation [14].

In chemical data analysis, the attention mechanism has shown promise in improving the performance of models by allowing for more fine-grained analysis of data [29]. For example, in gas chromatography data, the attention mechanism can be used to focus on specific peaks or retention times that are indicative of particular compounds or classes of compounds [11]. However, studies that incorporate existing attention mechanisms into various generative models for gas chromatography data show insignificant performance improvement results compared to models without attention mechanisms [11,30]. Therefore, we propose the importance of a new attention mechanism focusing on the inherent properties of gas chromatography data. The inherent properties of gas chromatography data are important for identifying and quantifying compounds such as retention time, mass-to-

charge ratio (m/z), peak area (PA), and peak height (PH) [31]. In this work, we propose a novel attention mechanism that focuses on the retention time among these properties.

The retention time, which is a characteristic of gas chromatography data, is a measure of the time it takes for a compound to pass through the chromatography column and reach the detector. The characteristic peaks that appear during these retention times are important characteristics of gas chromatography data because they can provide important information about the chemical composition of the sample.

Our proposed attention mechanism is used to emphasize the retention time feature in the chemical data and enable the generative model to produce more accurate and meaningful chemical data. The mechanism selectively focuses on retention time and its relationship with other chemical features, allowing the model to learn the complex dependencies and correlations that exist between these features. Therefore, the attention mechanism enables the generative model to produce more diverse and representative chemical data, which can be used to augment existing datasets and enhance the performance of chemical analysis and identification tasks.

3. Design Principle and Architecture

In this section, we propose an optimal preprocessing algorithm for gas chromatography data and *truncated* attention mechanism that efficiently learns gas chromatography data. Furthermore, we propose GCGAN as shown in Figure 1, a high-performance generative model that can conditionally generate gas chromatography data by applying various deep learning techniques, including transfer learning considering real chemical acquisition scenarios.

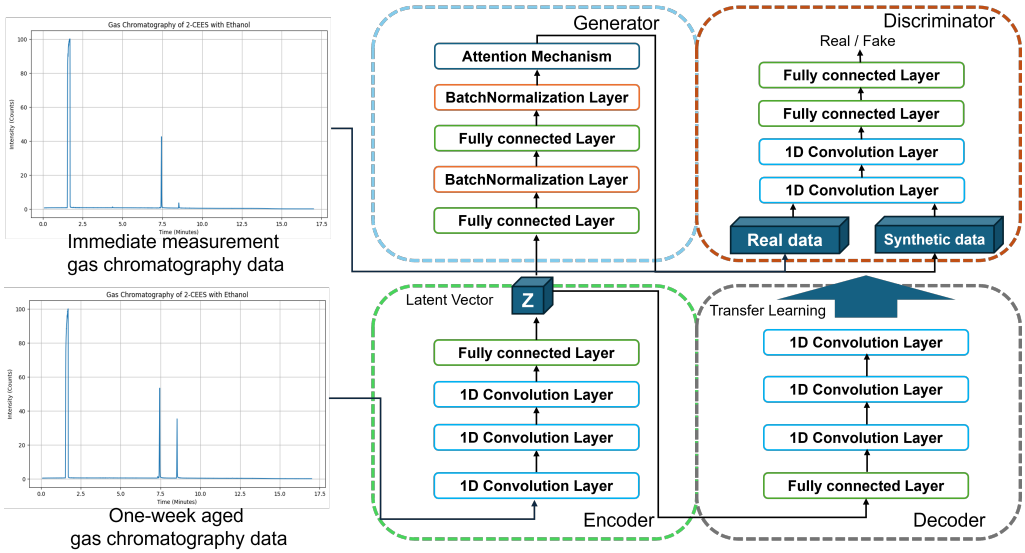


Figure 1. The overall structure of GCGAN.

3.1. Preprocessing Algorithm of GCGAN

We propose a preprocessing technique that allows the gas chromatography data in time-series format to be used appropriately in deep learning models such as the GCGAN proposed in this paper. The proposed preprocessing technique consists of sampling and robust scaling processes.

First, we sample complex forms of gas chromatography data consisting of large-scale time steps with a systematic sampling method that extracts them from the population with constant rules [32]. In addition to the large peak value at retention time, gas chromatography data are evenly distributed over most of the time, so the system sampling method is suitable, as shown in Figure 1, Ensuring that the samples selected during this sampling process are evenly distributed across the population can help reduce the risk of overfitting the original gas chromatography data for learning.

Additionally, we propose a method to use robust scaling for sampled data to allow deep learning models to properly learn outliers at the retention time of gas chromatography data. The robust scaler is a type of data normalization technique that scales the data based on the median and interquartile range (IQR) instead of the mean and standard deviation. The robust scaler used in this study is $x_{scaled} = \frac{x - \text{median}(X)}{\text{IQR}(X)}$, where x is the original value, median(X) is the median of the feature, and IQR(X) is the interquartile range of the feature. The interquartile range is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data [33]. Gas chromatography data often have extreme values, which can be outliers and may not follow a normal distribution, so robust scaling can help normalize the data and make them more suitable for analysis with machine learning algorithms [34]. Furthermore, robust scaling is less affected by the presence of outliers compared to other scaling methods such as $min - max$ scaling, which can make it more effective in preserving the information in the data.

3.2. Truncated Attention Mechanism

The *truncated* attention mechanism is designed to effectively capture and learn the unique characteristics of gas chromatography data, which consist of large peaks at retention time and small peaks at the rest of the time zone. The mechanism focuses on regions where the slope of the data changes rapidly, indicating the presence of peaks. Let $x(t)$ be the input data, where $t \in [0, T]$ represents the time step and T is the total number of time steps. The *truncated* attention mechanism can be formulated as follows in Equations (2)–(10). First, we define the slope function $s(t)$ as the absolute difference between adjacent time steps, as shown in Equation (2):

$$s(t) = |x(t) - x(t - h)| \tag{2}$$

where h is the time step size. Next, as shown in Equation (3), we introduce the attention weight function $a(t)$, which is determined by the slope function $s(t)$ and a threshold value τ:

$$a(t) = \sigma(\alpha \cdot (s(t) - \tau)) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, defined as shown in Equation (4):

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

The sigmoid function is a smooth, continuous function that maps any real-valued input to a value between 0 and 1. The threshold value τ is calculated as the mean of the slope function over the entire time range as shown in Equation (5):

$$\tau = \frac{1}{T} \int_0^T s(t)\, dt \tag{5}$$

The scaling factor α controls the steepness of the sigmoid function, determining how quickly the attention weight transitions from 0 to 1 around the threshold. When the slope function $s(t)$ is much larger than the threshold τ, the attention weight $a(t)$ approaches 1, indicating a strong emphasis on the corresponding time step. Conversely, when the slope function is much smaller than the threshold, the attention weight approaches 0, effectively truncating the influence of that time step.

The *truncated* attention mechanism can be applied to the input data $x(t)$ to obtain the attended data $\tilde{x}(t)$ as shown in Equation (6):

$$\tilde{x}(t) = a(t) \cdot x(t) \tag{6}$$

The attended data $\tilde{x}(t)$ preserve the large peaks while suppressing the small peaks and non-peak regions, effectively focusing the attention of model on the most informative parts of the data. The *truncated* attention mechanism can be further generalized to handle multidimensional input data $x(t) \in \mathbb{R}^n$, where n is the number of features. In this case, the

slope function and attention weight function can be applied element-wise to each feature as follows in Equations (7) and (8):

$$s_i(t) = |x_i(t) - x_i(t-h)|, \quad i = 1, 2, \ldots, n \tag{7}$$

$$a_i(t) = \sigma(\alpha \cdot (s_i(t) - \tau_i)), \quad i = 1, 2, \ldots, n \tag{8}$$

The threshold value τ_i is calculated as the mean of the slope function over the entire time range for each feature, as shown in Equation (9):

$$\tau_i = \frac{1}{T} \int_0^T s_i(t)\, dt \tag{9}$$

The multidimensional attended data $\tilde{x}(t)$ are obtained by element-wise multiplication of the attention weight functions and the input data, as shown in Equation (10):

$$\tilde{x}i(t) = ai(t) \cdot x_i(t), \quad i = 1, 2, \ldots, n \tag{10}$$

In contrast to existing attention mechanisms, our proposed *truncated* attention mechanism does not explicitly compute attention scores based on input sequences. Both approaches aim to focus on the critical parts of the input, but differ in how attention weights are calculated and applied to input sequences [28]. Existing attention mechanisms are to derive attention weights by calculating attention scores, but the proposed *truncated* attention mechanism serves the purpose of selectively focusing on specific parts of the input based on gradient and local variations.

By applying the *truncated* attention mechanism, the model can effectively focus on the regions with large peaks while suppressing the influence of small peaks and non-peak regions. This enables the model to capture the essential characteristics of the gas chromatography data and improve its learning performance.

The *truncated* attention mechanism provides a new method for processing data with different peak patterns, such as gas chromatography data, and has the potential to be applied to various domains in which similar data characteristics are observed. The mathematical formula presented above provides the basis for understanding and implementing the *truncated* attention mechanism in a deep learning model for gas chromatography data analysis.

3.3. Unified Structure of GCGAN

Various studies have been conducted to pretrain the generators of generative adversarial neural networks using autoencoders [35]. However, while only the latent vector of autoencoders is generally used for pretraining of GAN, we propose a unified learning scheme that uses autoencoders efficiently. Therefore, we unified two values of the autoencoder in GAN, as shown in Figure 2, to properly train gas chromatography. First, we use latent vectors with compressed data from encoders instead of using random variables for the generator in GCGAN. At this point, the generator of GCGAN begins meaningful mapping more proactively, and enables smooth learning when the sensitive information of the input gas chromatography learns compressed data to generate data. The latent vector of the autoencoder can help reduce the computational cost of the GAN training process, as it allows us to learn compressed representations of the data [36]. Furthermore, the use of latent vectors of the autoencoder can help overcome the mode decay problem in GAN training and ensure that the generator captures and generates important features of the input data [37].

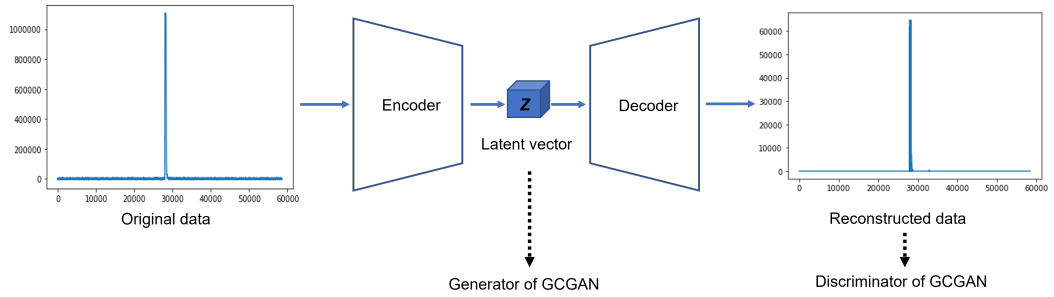
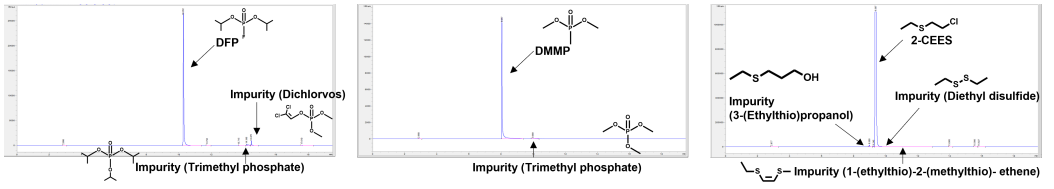
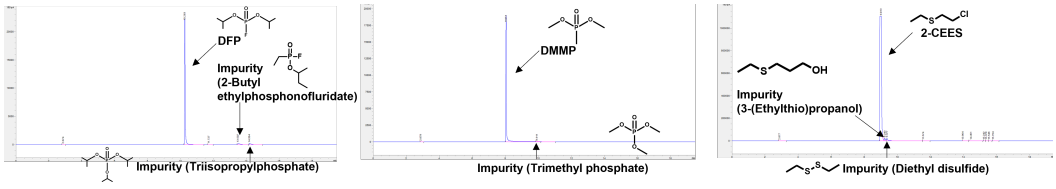


Figure 2. The unified structure of autoencoder in GCGAN.

Another part of our emphasis in GCGAN is the use of transfer learning from decoders, as shown in Figure 2. Transfer learning allows an artificial intelligence model to utilize the knowledge learned from one task for another related task [38]. For our proposed GCGAN, we apply transfer learning to discriminator networks that serve to distinguish between real and synthetic data. After acquiring a substance in an actual threatening situation scenario, various organic and chemical reactions may occur over time. Therefore, assuming this situation, we can learn various features and potential reactions of chemical data by pretraining a discriminator for chemicals that have reacted with solvents over a long period of time. In addition, the discriminator can quickly adapt to new tasks that distinguish real chemical data from synthetic chemical data without having to start a comparison task from scratch. The measurement results 1 week after mixing chemicals and solvent and mixing with those measured immediately are generally similar as shown in Figure 3, but the impurities are partially different. In the traditional GAN method, to learn and generate data that reacted immediately to the target, we repeatedly learn using random variables as latent vectors [12]. However, we propose efficient learning and using old data because it reuses aged data without discarding it. We leverage old data to ensure that our model aims at the robustness and generalization of synthetic data generation processes learned from a wider range of chemical reactions and their temporal progression. This approach not only preserves valuable data resources, but also enhances the ability of model to handle changes and anomalies in chemical data over time.



(**a**)



(**b**)

Figure 3. Gas chromatography data of various chemicals actually measured (**a**,**b**) show that various impurities were generated through chemical reactions in addition to the peak values at the retention time. (**a**) Data measured immediately after mixing chemicals with solvents. (**b**) Data measured 1 week after mixing the chemical with the solvent.

Overall, gas chromatography data do not significantly change the aspect of the data, as shown in Figure 3, even if a reaction occurs with a solvent over time. Therefore, using transfer learning on the discriminator has the following additional benefits:

- Pretrained discriminator for transfer learning can leverage the knowledge learned from related tasks to provide better discriminant performance, providing a more accurate and robust model.
- Transfer learning can help reduce overfitting, as pretrained models provide better normalization and can prevent the discriminator from remembering training data.
- Using transfer learning from discriminator leads to more efficient and effective models, providing better results and reducing the time and resources required for training.

3.4. Structure of GCGAN

We construct the encoder of GCGAN as a 1D CNN layer, allowing us to capture the time dependence of gas chromatography, as shown in Figure 1. This allows the generator of GCGAN to receive data in which the time information of gas chromatography is compressed, and to generate a sample that preserves the time pattern of the data.

The CNN is often used to process visual data, where shape and pattern are important, such as images [39]. For the data generated by the aforementioned preprocessing algorithm, GCGAN first uses a CNN to grasp the relationship and pattern between the environment of data with time-series characteristics [40]. Through this, the 1D CNN extracts important features from the data and generates a latent vector. The calculation process of the 1D CNN neural network output is as shown in Equation (11):

$$y_{i,k} = b_k + \sum_{j=0}^{F-1} w_{k,j} \cdot x_{i+j}. \tag{11}$$

In this formula, $y_{i,k}$ represents the k-th feature map at position i, b_k is the bias term for the k-th filter, $w_{k,j}$ is the weight for the j-th element of the k-th filter, and x_{i+j} is the input signal value at position $i + j$. The sum is taken over the filter length F, and the dot product between the filter weights and the input signal values is computed for each position i. When the result of a 1D CNN layer is compressed in an autoencoder, it is flattened into a 1D vector for potential representation of the input signal. Therefore, given the gas chromatography input signal x and the y of Equation (11), the latent representation z of the input signal can be calculated as $z = \text{flatten}(y)$. We use this latent vector z as an input to the generator network of GCGAN for proposed unified structure. Therefore, the generator function of GCGAN can be expressed as $G(\text{flatten}(y))$. The retention time t_r is incorporated into the latent vector z during the encoding process. This allows z to maintain the time structure of t_r and for the generator to generate data that accurately reflect the retention time characteristics. This relationship is denoted by $z = \text{flatten}(y_{t_r})$, where y_{t_r} encodes the retention time information. The discriminator network of GCGAN distinguishes real gas chromatography data x from synthetic data of the generator. As shown in Figure 2, the output reconstructed by the decoder of the autoencoder model for input gas chromatography data x is $f_{trans}(x)$. Therefore, we use the discriminator for input x to use the proposed transfer learning technique, as shown in Equation (12):

$$D(x; W_d) = \sigma(W_d \cdot f_{trans}(x)). \tag{12}$$

In the above Equation (12), we use σ as a sigmoid activation function, W_d as the weight matrix, and $\cdot$ as a matrix multiplication.

In GCGAN, our proposed *truncated* attention mechanism can be applied to the discriminator to improve its ability to distinguish between real and synthetic gas chromatography data. The *truncated* attention mechanism allows the discriminator to focus on data with most small values and only certain parts having very large values, which is particularly useful for the properties of gas chromatography data. The mathematical expression for the discriminator using the *truncated* attention mechanism is shown in Equation (13):

$$D(x) = \sigma\left(W_d \sum_{i=1}^{L} \alpha_i z_i + b_d\right). \tag{13}$$

In the above Equation (13), we use W_d and b_d as the weight matrix and bias term of the fully connected layer in the discriminator, z_i as the i-th feature map of the input data x obtained from the CNN layer, and α_i as the attention weight for the i-th feature map.

According to Equations (12) and (13), the discriminator model of GCGAN that we propose is as shown in Equation (14):

$$D(X; W_d) = \sigma\left(W_d \cdot f_{trans}(X) \sum_{i=1}^{L} \alpha_i z_i + b_d\right). \tag{14}$$

In GCGAN, $G(\text{flatten}(y))$ is trained to deceive $D(X; W_d)$ by generating generated data that are difficult to distinguish from real data. The objective of $G(\text{flatten}(y))$ is to maximize the error rate of $D(X; W_d)$, while the objective of $D(X; W_d)$ is to minimize its error rate as in Equation (1).

4. Performance Evaluation

In this section, we describe our experiments to augment chemical data using GCGAN with *truncated* attention mechanism and transfer learning technique. Therefore, we implement all of the proposed algorithms and measure their performance through the measured experimental chemical data.

4.1. Datasets

We used data that experimentally measured the gas chromatography of DMMP, DFP, and 2-CEES using the Agilent 8890 GC system (G3540A). Each chemical was measured under the experimental conditions as shown in Table 1 in the device such as Figure 4c, and each chemical was measured under the experimental conditions, as shown in Table 2, and we proceeded based on the standard experimental protocol in the experiment [41].



(**a**) (**b**) (**c**)

Figure 4. Experimental acquisition process of gas chromatography (GC) data. (**a**) Process of combining chemicals and solvents. (**b**) Process of mixing chemicals and solvents. (**c**) Analyze with GC equipment.

Table 2. Experimental conditions of gas chromatography (GC) data.

GC Conditions	
Injection source	Manual
Inlet	Splitless, 200 °C
Column	HP-5MS 5 % Phenyl Methyl Silox. 30 m × 250 µm × 0.25 µm
Provider of GC column	Agilent Technologies, Santa Clara, CA, USA
Oven Temperature Program	40 °C (3 min) 10 °C/min to 120 °C (13 min) 30 °C/min to 240 °C (17 min)
Injection volume	1 µL
Gas flow conditions	
Split Ratio	30:1
Split Flow	30 mL/min
Column Flow	1.1 mL/min

DMMP, DFP, and 2-CEES are all chemicals that are commonly used as simulant agents in the testing and development of chemical sensors and detectors [42]. The purity of these substances with ethanol solvent is as follows: 2-CEES (97%, Chemical Abstract Service (CAS) Number 693-07-2), DMMP (>98%, CAS Number 756-79-6), and DFP (>98%, CAS Number 55-91-4).

- DMMP (dimethyl methylphosphonate) is a colorless, odorless, and highly volatile liquid that is used as a simulant for nerve agents such as sarin and soman. DMMP is structurally similar to these nerve agents, making it a useful tool to test the effectiveness of sensors and detectors designed to detect these dangerous compounds [43].
- DFP (diisopropyl fluorophosphate) is a potent organophosphorus compound that is used as a simulant for nerve agents. DFP is highly toxic and is commonly used in research laboratories as a model compound to study the effects of nerve agents on living organisms. It is also used as a standard reference material for chemical analysis [44].
- 2-CEES (2-chloroethyl ethyl sulfide) is a simulant of mustard gas, a chemical warfare agent that causes severe blistering of the skin and damage to the respiratory system. 2-CEES is structurally similar to mustard gas and is often used in the testing of chemical detectors and protective equipment designed to detect or mitigate exposure to this dangerous compound [45].

Each substance consists of a pair of data detected immediately on the day and measured after a chemical reaction for a week, as shown in Figure 3. Solvent is used for moving each material in a gas chromatography device. We measured each chemical by mixing it with an ultrasonic wave detector with solvents as shown in Figure 4b.

Therefore, we experimentally confirmed that impurities such as ethyl dipentyl phosphate, triisopropyl phosphate, etc. are produced through the reaction with solvent for 1 week, as shown in Figure 3b. The temporal spacing between these data is a scenario that assumes when a real scenario of situations where measurement is limited, such as chemical terrorism occurs, and we generate synthetic gas chromatography data by transferring gas chromatography data over time for reasons such as transport of materials.

4.2. Implementation

We used an Intel Dual-Core Xeon CPU @ 2.30GHz, 32 GB RAM, Nvidia Tesla P100 GPU for training of full implemented GCGAN and gas chromatography data augmentation. Our GCGAN processes GC data time series to effectively capture and model retention time properties in chromatography columns. To implement the proposed GCGAN, we set the hyperparameters of the part of the autoencoder and the part of GAN that make up the GCGAN, as shown in Table 3.

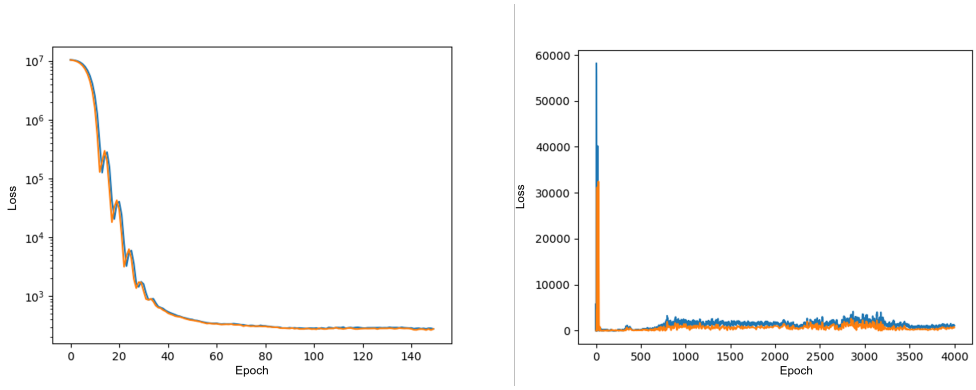
Table 3. Hyperparameters used in GCGAN.

(**a**) **Hyperparameters used in autoencoder**	
Hyperparameter	**Value**
Learning rate	0.0004
Loss	MSE
Epoch	150
Optimizer	Adam
Early stopping patience	50
(**b**) **Hyperparameters used in GAN**	
Hyperparameter	**Value**
Learning rate	0.0007
Loss	BCE
Epoch	1000
Optimizer	Adam
Early stopping patience	50

The batch size is trained as a whole for sampling by the preprocessing algorithm proposed in Section 3. Mean squared error (MSE) loss used in Table 3a is a loss function that measures the mean square difference between the predicted value and the true value. MSE loss is used to calculate the difference between the values of the result of the autoencoder and the actual chromatography data. MSE loss measures the difference between binary values of input and output, indicating that the smaller the loss, the better the reconstruction. This makes it a suitable loss function for the autoencoder for gas chromatography data, which aims to minimize reconstruction errors between input and output. On the other hand, binary cross-entropy (BCE) loss used in Table 3b measures the discrepancy between the predicted probability distribution and the actual probability distribution. The generator is then updated to minimize this difference so that the generated data are more similar to the actual data. By minimizing BCE loss, the generator is encouraged to generate gas chromatography data similar to the real data.

We implement an early stopping algorithm for optimized GCGAN training periods and stopping points to effectively prevent fitting. The algorithm monitors the root mean squared error (RMSE) between real and generated synthetic data in both the training and validation datasets and compares it with the best observed RMSE. The process meets the early stop criteria and stops training as soon as no improvement in RMSE is observed during 50 consecutive epochs according to predefined patience parameters such as Table 3b.

Consequently, we demonstrate that stable learning is possible, as shown in Figure 5, by the optimized implementation of the proposed GCGAN using gas chromatography dataset. In particular, in the GAN training process in Figure 5b, loss is rapidly reduced in the initial epoch. This indicates that our proposed unified structure of GCGAN and the *truncated* attention mechanism worked properly and learned efficiently.



(**a**) (**b**)

Figure 5. Loss of GCGAN in training process. The blue line represents training loss, and the orange line represents validation loss. (**a**) Loss of autoencoder in training process. (**b**) Loss of GAN in training process.

4.3. Evaluation Metrics

We use various metrics as follow to compare and evaluate the gas chromatography data generated by GCGAN with the original data. In addition, we implement a deep learning-based artificial intelligence model to demonstrate the superiority of synthetic gas chromatography data that are difficult to represent with existing evaluation metrics.

- Visual inspection: We propose a method for visually inspecting generated gas chromatography data using the original data and graphs to demonstrate the performance of GCGAN. It is a method to display both the original data and the generated data on the same graph and to compare the values of retention time and peak value, which are important indicators for gas chromatography, as mentioned in Section 2. Since each chemical has its own retention time, this method shows that GCGAN performs well if the generated data are very similar to the original data [46].

- Quantitative evaluation: We use the Pearson correlation coefficient (PCC), the Spearman correlation coefficient (SCC), and cosine similarity techniques to quantitatively evaluate the performance of GCGAN. PCC, SCC, and cosine similarity are commonly used metrics to measure chromatographic similarity. To proceed with the accurate evaluation, we generate 10 synthetic data for each data and use the data obtained by averaging the peak value for each timestamp of the generated data. PCC is one of the metrics appropriately used to evaluate the performance of machine learning models, including GAN, in various applications [47]. SCC is a nonparametric measure of the monotonicity of the relationship between two datasets, which can be used to assess the similarity of the original and generated gas chromatography data, even if their relationship is not linear [48]. Cosine similarity measures the cosine of the angle between two vectors, providing a value between −1 and 1, with 1 indicating high similarity, 0 indicating no similarity, and −1 indicating dissimilarity. In the context of reinforcing gas chromatography data with GAN, PCC, SCC, and cosine similarity can be used to compare the original data with the generated data and evaluate the similarity between the two datasets. PCC is a measure of the linear correlation between two variables and ranges from −1 to 1, with 1 indicating a perfect positive correlation, 0 indicating no correlation, and −1 indicating a perfect negative correlation. By using these three metrics together, we can comprehensively assess the similarity between the original and generated gas chromatography data, ensuring the quality and reliability of the GCGAN-augmented data [49].
- Using deep learning model: To demonstrate the usefulness of chemical synthetic data, we also build a very basic and primitive discriminant model consisting of a single dense layer. This simple model is designed to show the effect of progressively learning from the synthetic data generated by GCGAN. Our evaluation is experimented by using a deep neural network (DNN) with a fully connected layer. The architecture was developed by a backpropagation neural network (BPNN) model used for research dealing with complex patterns inherent in chromatographic data, which is used to effectively evaluate the quality of synthetic data generated by GCGAN and to evaluate the usefulness of synthetic data [18]. This experiment highlights the effectiveness of GCGAN-generated data in improving the learning ability of the most basic discriminant model. By gradually incorporating synthetic data, we observe a noticeable increase in the accuracy of the model, emphasizing the value of synthetic data in improving the performance of gas chromatography data classification tasks even in rudimentary model architectures. One of the contributions of this study is to improve the performance of artificial intelligence models through high-quality synthetic data generation. The deep learning model has the characteristic that the performance improves as the number of appropriate data required for learning increases [50]. Therefore, we propose a method to implement a deep learning model based on fully connected layers for classification to measure the performance of the model according to the amount of learning of synthetic data. We measure the performance according to the amount of learning of synthetic data in two situations for accurate experiments of classification models on gas chromatography data. We generated 1, 15, and 50 synthetic data for each chemical for experiments that demonstrate the effectiveness of synthetic data. Each dataset instance, whether real or synthetic, follows a uniform shape of (2, 58,500), ensuring consistency in the data representation in all experiments. We used models trained on these datasets to measure the impact of data augmentation on classification performance using both synthetic and real data to evaluate the ability of the model to classify classes of three chemicals: 2-CEES, DFP, and DMMP. This approach effectively mitigates the risk of discriminator overfitting, which is generally a concern when training GAN with limited datasets [51].
  First, we conducted a classification experiment by generating a number of validation datasets using random variables drawn from a normal distribution, based on the

gas chromatography data of DMMP, DFP and 2-CEES, which were used to train the GCGAN.

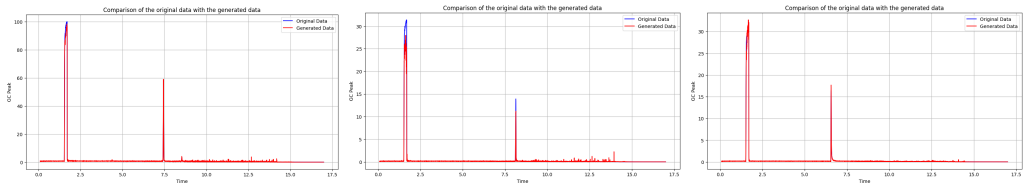
Second, we performed a more complex classification experiment by enriching the validation dataset from the first experiment with gas chromatography data not used in GCGAN training. For this, we include data of 2-Chloroethyl phenyl sulfide (2-CEPS), a chemical not previously involved in GCGAN training, with measurements from four solvents: ethanol, methanol, dimethyl carbonate, and tetrahydrofuran. Additionally, we incorporate data from a 1-week reaction of 2-CEES with methanol solvent into the validation dataset, providing further dimensions for model evaluation.

We measured the performance of deep learning models for chemical classification using confusion matrix-based accuracy and area under the receiver operating characteristic curve (AUC). The confusion matrix is a performance evaluation metric for a classification model and consists of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [52]. Accuracy is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, where a higher value indicates how many correct classifications were made during the entire classification. However, accuracy is not reliable when the data are imbalanced, which is the goal of classification. Additionally, we also measure AUC to evaluate the more accurate performance of the classification model on various gas chromatography data. AUC is the area under the receiver operating characteristic curve that represents the change in the TP rate as $\frac{TP}{TP+FN}$ and the FP rate as $\frac{TN}{FP+TN}$ [53]. As a result, we demonstrated the performance of GCGAN by showing that the more synthetic data we have, the better the classification model.

4.4. Evaluation Results

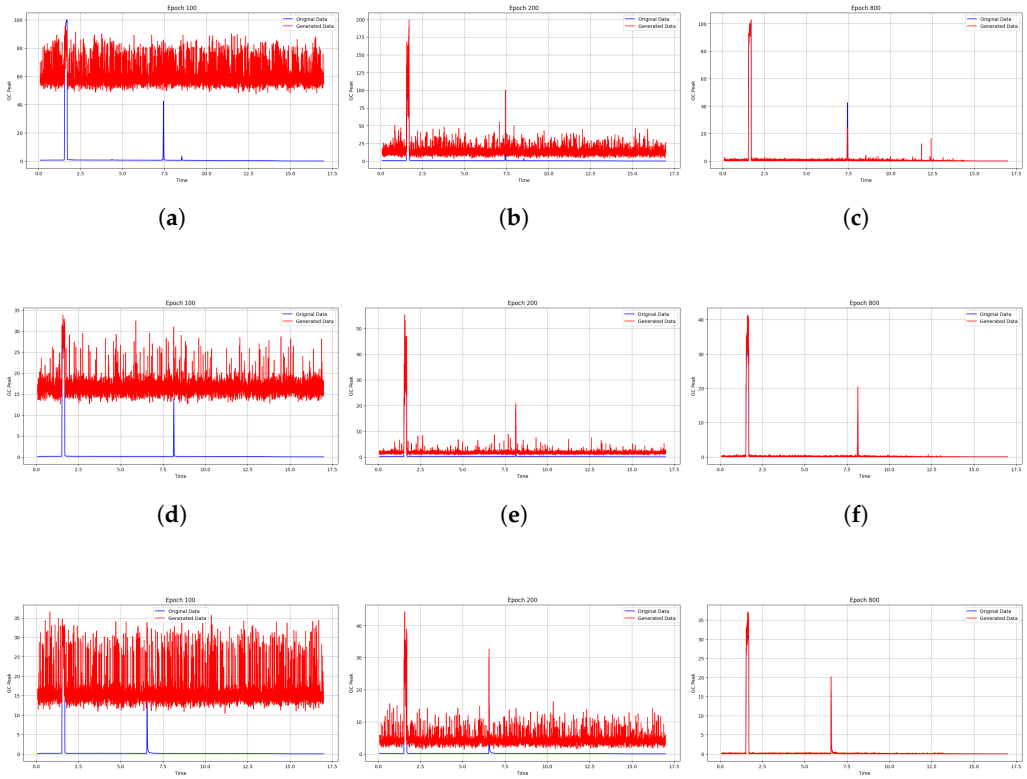
We successfully conditionally generated synthetic data for each of the three substances using GCGAN, as shown in Figure 6. These results indicate that our GCGAN was able to effectively augment the chemical data and capture the important features of the input data. We experimented with training up to 4000 epochs, as shown in Figure 5b, to find a suitable epoch, and confirmed that the loss of GCGAN quickly stabilizes before 1000 epochs. Therefore, we set the hyperparameters as in Table 3. The GCGAN set through this indicates that the learning process of 2-CEES, DFP, and DMMP all proceeds appropriately, as shown in Figure 7.

Visual inspection We show the comparison of the synthetic data generated by GCGAN with the original data for three different chemicals—2-CEES, DFP, and DMMP—through Figure 6. The blue line represents the original data for each material and the red line represents the corresponding synthetic data. It is clear that the retention time and peak values of the data generated by visual inspection agree very well with the retention time and peak values of the original data for all three chemicals. It shows the effectiveness of GCGAN in capturing the intrinsic properties of gas chromatography data and generating realistic synthetic samples that can be used for data augmentation purposes. Furthermore, we conducted experiments to demonstrate the effectiveness of the *truncated* attention mechanism. We trained 2-CEES by implementing GAN with the same structure and hyperparameter of GCGAN as proposed in the paper, but without the *truncated* attention mechanism, and proved that learning is difficult until 1000 epochs, as shown in Figure 8. This suggests that the *truncated* attention mechanism is suitable for time-series data with large deviations, such as gas chromatography.



(**a**) (**b**) (**c**)

Figure 6. Comparison of the synthetic data with the original data. The blue line represents the original data for each material and the red line is the synthetic data. In each graph, the horizontal axis represents time (minutes) and the vertical axis represents the peak value. (**a**) Comparison of the synthetic data with the original data about 2-CEES. (**b**) Comparison of the synthetic data with the original data about DFP. (**c**) Comparison of the synthetic data with the original data about DMMP.

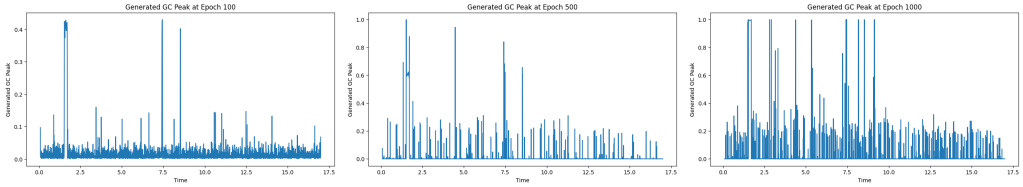


(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)

Figure 7. Process of gas chromatography data generation. The blue line is the original data of each material, and the red line is the synthetic data generated during the training process. In each graph, the horizontal axis represents time (minutes), and the vertical axis represents the peak value. (**a**) Training of 2-CEES at 100 epochs. (**b**) Training of 2-CEES at 200 epochs. (**c**) Training of 2-CEES at 800 epochs. (**d**) Training of DFP at 100 epochs. (**e**) Training of DFP at 200 epochs. (**f**) Training of DFP at 800 epochs. (**g**) Training of DMMP at 100 epochs. (**h**) Training of DMMP at 200 epochs. (**i**) Training of DMMP at 800 epochs.



(**a**) (**b**) (**c**)

Figure 8. Results of training GCGAN without *truncated* attention mechanism. In each graph, the horizontal axis represents time (minutes), and the vertical axis represents the peak value. (**a**) Training of 2-CEES at 100 epochs without *truncated* attention mechanism. (**b**) Training of 2-CEES at 500 epochs without *truncated* attention mechanism. (**c**) Result of 2-CEES at 1000 epochs without *truncated* attention mechanism.

Quantitative evaluation We show in Table 4 that synthetic data were generated similar to the chemical properties of the original data. The table presents the quantitative evaluation results of the synthetic data generated for three different chemicals: 2-CEES, DFP, and DMMP. We can see that the PCC values for the three chemicals are extremely high, ranging from 0.9965 to 0.9984, indicating a strong linear correlation between the original and generated data [54]. The SCC values, which measure the monotonic relationship between the datasets, are also relatively high, ranging from 0.8192 to 0.8352, suggesting a strong overall similarity in the rank order of the data points. Finally, the cosine similarity values are identical to the PCC values, further confirming the high degree of similarity between the original and generated data vectors. Overall, we represent the effectiveness of the GCGAN model in generating synthetic gas chromatography data that closely mimic the characteristics of the original data for various chemicals [49]. The high values of PCC, SCC, and cosine similarity demonstrate that the generated data capture the essential features and patterns present in the original data, validating the quality and reliability of the synthetic data generated by GCGAN for the enhancement of gas chromatography datasets.

Table 4. Quantitative evaluation of generated synthetic data.

Chemical	PCC	SCC	Cosine Similarity
2-CEES	0.9976	0.8352	0.9976
DFP	0.9965	0.8192	0.9965
DMMP	0.9984	0.8322	0.9984

Additionally, we perform validation based on more additional chemical analysis methods to further validate the quality of the synthetic data generated by the proposed model as shown below [55,56]:

- Peak area (PA): The peak area is calculated as the sum of the values under a peak, which represents the concentration of the compound.
- Peak height (PH): The peak height is the maximum value of the peak, which also correlates with the concentration of the compound.

The similarity verification of the generated synthetic data in terms of chemical analysis indicates the quality of the synthetic data through the PA and PH, as shown in Table 5. Specifically, the PA of DMMP, 2-CEES, and DFP show similarities of 98.31%, 97.24%, and 97.10%, respectively, indicating high accuracy in the concentration expression of the synthetic data. Similarly, the PH of DMMP, 2-CEES, and DFP show similarities of 69.20%, 87.28%, and 97.44%, respectively. Although these results show the need for room for improvement in 2-CEES of PH, they show similarities in excess of 90% overall in terms of peak-based chemical analysis. This verifies the effectiveness of the proposed data augmentation method by emphasizing that the synthetic data retain the characteristics of the original data.

Table 5. Evaluate the chemical analysis similarity of the generated synthetic data.

Metric	Real DMMP	Synthetic DMMP	Real 2-CEES	Synthetic 2-CEES	Real DFP	Synthetic DFP
Peak Area	6.8863	7.0025	24.2675	23.5981	6.76681	6.9633
Peak Height	30.8101	40.301	100.00	112.7194	31.4042	32.2071

Using deep learning model We demonstrate that synthetic gas chromatography data improve the performance of classification models, as shown in Table 6.

Table 6. Experiments on performance improvement of classification model. (**a**) shows the experimental results of the classification model for the validation dataset consisting of DMMP, DFP, and 2-CEES, and (**b**) shows the experimental results of the classification model for the dataset with 2-CEPS for various solvents and various time conditions.

(**a**) **Experiment on Validation GC Data**			
Size of Data	**1**	**15**	**50**
Accuracy	0.3367	0.6667	1
AUC	0.5025	0.6944	1
(**b**) **Experiment GC Data on Various Conditions**			
Size of Data	**1**	**15**	**50**
Accuracy	0.3067	0.6667	0.8133
AUC	0.5485	0.8036	0.9357

Table 6 shows the performance results of the classification model that was trained using synthetic data after generating 1, 15, and 50 synthetic data using GCGAN for each of the three chemicals. This experiment crucially represents that the ability of the model to identify mixed chemical data significantly improves with a much larger performance observed with 50 augmentation per datum, with even a slight augmentation with 15 per real datum.

We highlight a very important contribution to improving the performance of the urgent identification model for hazardous chemicals used in this paper.

In the first experiment using only validation data, the classification model achieved an accuracy of 0.3367 and an AUC of 0.5025 when trained on only 1 pair of training datasets as shown in Table 6a. This baseline performance establishes the initial capability of our model to classify gas chromatography data without the aid of augmented data. When the model is subsequently trained with increasing quantities of augmented data, the accuracy and AUC results gradually increased, demonstrating the added value of synthetic data in enhancing the model. Accuracy and AUC achieved a very high value of 1 when trained with 50 synthetic data, indicating that the inclusion of diverse synthetic samples significantly enriches the training set, leading to perfect classification performance.

In the second experiment using untrained data, the classification model achieved an accuracy of 0.3067 and an AUC of 0.5485 when trained on only one pair of training datasets, as shown in Table 6b. Similar to the first experiment, this initial result highlights the challenges faced by the classification model when limited to sparse training data. By training with increasing quantities of augmented data up to 50, we achieved AUC and accuracy of 0.8133 and 0.9357, respectively. The substantial improvement in both metrics underscores the effectiveness of synthetic data in providing the model with a richer, more comprehensive understanding of the data distribution. This represents a significant improvement over the model trained on the original data alone.

Additionally, we conduct experiments as shown in Figure 9 to demonstrate the effectiveness of using vectors extracted from encoder of autoencoder in GCGAN as shown in Figure 2. To this end, we use the data generated by *truncated* attention GAN used based on normal distribution random variables, as in conventional GAN training methods, and the data generated by GCGAN of the complete structure we propose. As shown in Figure 9, when the data size is 1, identifying the data of DMMP, DFP, and 2-CEES and classifying

the correct label shows low numbers in both cases. However, it can be seen that the more synthetic data generated for learning the classification model, the more the GCGAN of the complete structure that we propose can improve performance rapidly. This improvement is remarkable as it progresses from 15 to 50 datasets, demonstrating the robustness of the generated synthetic data in enriching the training environment.

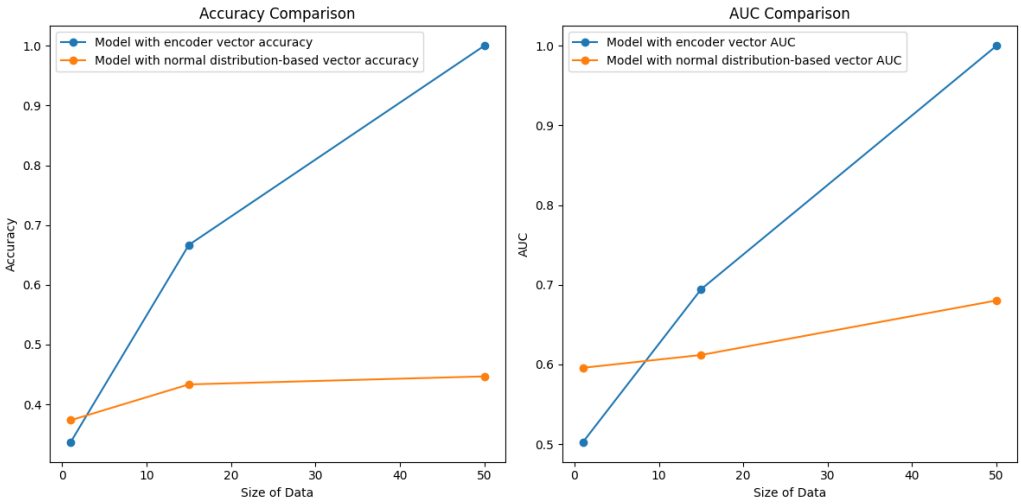


Figure 9. Effectiveness of using encoder vectors in GCGAN. The blue line is a method using a vector extracted from the encoder we propose, and the orange line is a method using a normal distribution-based vector instead of an encoder vector.

The results in Figure 9 not only demonstrate the efficiency of the proposed GCGAN structure, but also highlight the efficiency of using encoder extraction vectors over conventional random sampling methods. This implies that the GCGAN method can use the intrinsic properties of the input data more effectively, which is essential for generating high-fidelity synthetic data. These results indicate that the method of generating using vectors from the encoder is effective and suggest that our method can make a significant contribution to various AI models used in the chemical field.

Overall, our results demonstrate that augmenting gas chromatography data with synthetic data generated by GCGAN can improve the performance of an artificial intelligence classification model. It is particularly noteworthy that the synthetic data not only complements but also significantly extends the representational diversity of the training set, thereby enhancing the predictive accuracy and generalization capability of model. Furthermore, experimental results confirm that the GCGAN can produce high-quality synthetic data that closely resemble the original data, validating our approach as a viable method for data augmentation in gas chromatography analysis.

4.5. Discussion

Our results demonstrate that GCGAN with *truncated* attention mechanism and transfer learning technique in unified model can effectively augment chemical data and improve the performance of chemical modeling. The use of *truncated* attention mechanism and the transfer learning technique improved the ability of discriminator to distinguish between real and generated data. Furthermore, we found that GCGAN configured without *truncated* attention mechanism and transfer learning during implementation cannot learn gas chromatography data no matter how advanced the neural network is. Therefore, our *truncated* attention mechanism and transfer learning technique in the unified model are suitable for learning time-series data such as gas chromatography data, where the deviation between the value at a certain time and the value at the other time is very large.

Our approach has several potential applications in drug discovery and chemical modeling. By augmenting chemical data, we can generate larger datasets that can be used for training machine learning models. This can improve the accuracy and reliability of

these models and help accelerate the drug discovery process. Our approach can also be used to generate novel compounds that may have potential therapeutic properties [57].

However, there are some limitations to our approach that need to be addressed in future work. One limitation is that the generated chemical data may not always be chemically realistic. Future research may explore ways to ensure that the compounds produced are physically and chemically realistic. In addition, while the used metrics may not fully capture the complexity of the problem. Future work can explore alternative metrics that are more suitable for the task of generating chemical data. Furthermore, we plan to apply this attention mechanism to other important parameters, such as m/z in GC data, to improve the robustness and applicability of the generative model in future work. We also aim to validate our model on chemicals measured under various time and mixed conditions to ensure comprehensive performance evaluations, along with metrics such as resolution and contrast.

5. Conclusions

Through this study, we represent a generative adversarial neural network based data augmentation technique that can be applied to time-series data of chemicals. In addition, this algorithm, which generates simulated data similar to time-series data measured with gas detection equipment, such as gas chromatography analysis for chemicals, can also generate simulated data for chemicals whose individual properties change over time, such as gas chromatography.

We summarized the contents and contributions of this study as follows:

- We develop an novel attention mechanism that pays attention to a specific critical portion of all of the data in gas chromatography data.
- We designed GCGAN with transition learning for scenarios that take into account actual chemical acquisition time, and we demonstrated performance by implementing them all in practice.
- We demonstrated the performance of GCGAN by using directly experimentally acquired gas chromatography data, not open-source or simulation data, for the performance evaluation of GCGAN.

To improve the proposed study, we plan future studies as follows:

- We have implemented a classification model consisting of deep learning layers to evaluate the quality of synthetic data generated by GCGAN. Our further work could lead to the development of a more advanced gas chromatography data classification model using a novel attention mechanism or neural network layers, which is very necessary in the field of chemical analysis.
- Future research should explore how to visualize and analyze the generated compounds to better understand the performance of the model.

The simulated data generated in this way can be linked to research that develops toxic chemical detection algorithms and improves performance through identifying singularities and learning patterns. In the future, research is planned to increase the diversity of simulated data generation by applying various statistical techniques to actual data in the preprocessing process.

Author Contributions: Conceptualization, N.Y. and H.K.; methodology, N.Y. and H.K.; software, N.Y. and W.J.; validation, N.Y. and W.J.; formal analysis, N.Y. and H.K.; investigation, N.Y. and H.K.; resources, N.Y. and H.K.; data curation, N.Y. and H.K.; writing—draft preparation, N.Y. and W.J.; writing—review and editing, N.Y. and W.J.; visualization, N.Y. and W.J.; supervision, H.K.; project administration, H.K.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Research Institute of defense Technology planning and advancement (KRIT) grant funded by Defense Acquisition Program Administration (DAPA) (KRIT-CT-21-034).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly provided under the terms of the contract with the funding agency.

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

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative adversarial networks
AE	Autoencoder
DL	Deep learning
GCGAN	Gas chromatography GAN
SMILES	Simplified molecular input line entry system
CNN	Convolutional neural networks
G	Generator
D	Discriminator
IQR	Interquartile range
DMMP	Dimethyl methylphosphonate
DFP	Diisopropyl fluorophosphate
2-CEES	2-Chloroethyl Ethyl Sulfide
MSE	Mean squared error
BCE	Binary cross entropy
RMSE	Root mean squared error
AUC	Area under receiver operating characteristic curve
TP	True positive
TN	True negative
FP	False positive
FN	False negative

References

1. Ciottone, G.R. Toxidrome recognition in chemical-weapons attacks. *N. Engl. J. Med.* **2018**, *378*, 1611–1620. [CrossRef]
2. Greenfield, R.A.; Slater, L.N.; Bronze, M.S.; Brown, B.R.; Jackson, R.; Iandolo, J.J.; Hutchins, J.B. Microbiological, biological, and chemical weapons of warfare and terrorism. *Am. J. Med. Sci.* **2002**, *323*, 326–340. [CrossRef]
3. Valdez, C.A.; Leif, R.N.; Hok, S.; Hart, B.R. Analysis of chemical warfare agents by gas chromatography-mass spectrometry: Methods for their direct detection and derivatization approaches for the analysis of their degradation products. *Rev. Anal. Chem.* **2017**, *37*, 20170007. [CrossRef]
4. Krone, N.; Hughes, B.A.; Lavery, G.G.; Stewart, P.M.; Arlt, W.; Shackleton, C.H. Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *J. Steroid Biochem. Mol. Biol.* **2010**, *121*, 496–504. [CrossRef]
5. Lee, W.; Lee, J.Y.; Kim, H. Mobile device-centric approach for identifying problem spot in network using deep learning. *J. Commun. Netw.* **2020**, *22*, 259–268. [CrossRef]
6. Joo, H.; Lee, S.; Lee, S.; Kim, H. Optimizing Time-Sensitive Software-Defined Wireless Networks with Reinforcement Learning. *IEEE Access* **2022**, *10*, 119496–119505. [CrossRef]
7. Yoon, N.; Jung, W.; Kim, H. DeepRSSI: Generative Model for Fingerprint-Based Localization. *IEEE Access* **2024**, *12*, 66196–66213. [CrossRef]
8. Baum, Z.J.; Yu, X.; Ayala, P.Y.; Zhao, Y.; Watkins, S.P.; Zhou, Q. Artificial intelligence in chemistry: Current trends and future directions. *J. Chem. Inf. Model.* **2021**, *61*, 3197–3212. [CrossRef]
9. Risum, A.B.; Bro, R. Using deep learning to evaluate peaks in chromatographic data. *Talanta* **2019**, *204*, 255–260. [CrossRef]
10. Baccolo, G.; Quintanilla-Casas, B.; Vichi, S.; Augustijn, D.; Bro, R. From untargeted chemical profiling to peak tables–A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends Anal. Chem.* **2021**, *145*, 116451. [CrossRef]
11. Yoon, N.; Kim, H. Pioneering AI in Chemical Data: New Frontline with GC-MS Generation. In Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 19–22 February 2024; p. 826.
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]

13. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
14. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
15. Liu, B.; Zhu, Y.; Song, K.; Elgammal, A. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
16. Malmquist, G.; Danielsson, R. Alignment of chromatographic profiles for principal component analysis: A prerequisite for fingerprinting methods. *J. Chromatogr. A* **1994**, *687*, 71–88. [CrossRef]
17. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48.
18. Wang, Y.; He, T.; Wang, J.; Wang, L.; Ren, X.; He, S.; Liu, X.; Dong, Y.; Ma, J.; Song, R.; et al. High performance liquid chromatography fingerprint and headspace gas chromatography-mass spectrometry combined with chemometrics for the species authentication of Curcumae Rhizoma. *J. Pharm. Biomed. Anal.* **2021**, *202*, 114144. [CrossRef] [PubMed]
19. Matyushin, D.D.; Sholokhova, A.Y.; Buryak, A.K. Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases. *Int. J. Mol. Sci.* **2021**, *22*, 9194. [CrossRef]
20. Vaškevičius, M.; Kapočiūtė-Dzikienė, J.; Šlepikas, L. Prediction of chromatography conditions for purification in organic synthesis using deep learning. *Molecules* **2021**, *26*, 2474. [CrossRef] [PubMed]
21. Fedorova, E.S.; Matyushin, D.D.; Plyushchenko, I.V.; Stavrianidi, A.N.; Buryak, A.K. Deep learning for retention time prediction in reversed-phase liquid chromatography. *J. Chromatogr. A* **2022**, *1664*, 462792. [CrossRef]
22. Vrzal, T.; Malečková, M.; Olšovská, J. DeepReI: Deep learning-based gas chromatographic retention index predictor. *Anal. Chim. Acta* **2021**, *1147*, 64–71. [CrossRef]
23. Chan, S.; Santoro, A.; Lampinen, A.; Wang, J.; Singh, A.; Richemond, P.; McClelland, J.; Hill, F. Data distributional properties drive emergent in-context learning in transformers. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18878–18891.
24. Kim, S.; Noh, J.; Gu, G.H.; Aspuru-Guzik, A.; Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS Cent. Sci.* **2020**, *6*, 1412–1420. [CrossRef] [PubMed]
25. Hussein, S.A.; Tirer, T.; Giryes, R. Image-adaptive GAN based reconstruction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 3121–3129.
26. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
27. Han, Y.; Fan, C.; Xu, M.; Geng, Z.; Zhong, Y. Production capacity analysis and energy saving of complex chemical processes using LSTM based on attention mechanism. *Appl. Therm. Eng.* **2019**, *160*, 114072. [CrossRef]
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* , 5998–6008.
29. Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914–923. [CrossRef] [PubMed]
30. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
31. Sánchez-Guijo, A.; Hartmann, M.F.; Wudy, S.A. Introduction to gas chromatography-mass spectrometry. In *Hormone Assays in Biological Fluids*; Humana Press: Totowa, NJ, USA, 2013; pp. 27–44.
32. Yates, F. Systematic sampling. *Philos. Trans. R. Soc. Lond. Ser. Math. Phys. Sci.* **1948**, *241*, 345–377.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Ahsan, M.M.; Mahmud, M.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* **2021**, *9*, 52. [CrossRef]
35. Ham, H.; Jun, T.J.; Kim, D. Unbalanced gans: Pre-training the generator of generative adversarial network using variational autoencoder. *arXiv* **2020**, arXiv:2002.02112.
36. Ma, R.; Lou, J.; Li, P.; Gao, J. Reconstruction of generative adversarial networks in cross modal image generation with canonical polyadic decomposition. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 8868781. [CrossRef]
37. Tran, N.T.; Bui, T.A.; Cheung, N.M. Dist-gan: An improved gan using distance constraints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–385.
38. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
39. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
40. Srinivasamurthy, R.S. Understanding 1D Convolutional Neural Networks Using Multiclass Time-Varying Signalss. Ph.D. Thesis, Clemson University, Clemson, SC, USA, 2018.
41. Agilent Technologies. *User Manual for GC Maintenance 8890*. Agilent Technologies: Santa Clara, CA, USA, 2020. Available online: https://www.agilent.com/cs/library/usermanuals/public/usermanual-gc-maintenance-8890-g3540-90015-en-agilent.pdf (accessed on 1 November 2020).
42. Bartelt-Hunt, S.L.; Knappe, D.R.; Barlaz, M.A. A review of chemical warfare agent simulants for the study of environmental behavior. *Crit. Rev. Environ. Sci. Technol.* **2008**, *38*, 112–136. [CrossRef]

43. Xiang, H.; Xu, H.; Wang, Z.; Chen, C. Dimethyl methylphosphonate (DMMP) as an efficient flame retardant additive for the lithium-ion battery electrolytes. *J. Power Sources* **2007**, *173*, 562–564. [CrossRef]
44. Leopold, I.H.; Comroe, J.H. Effect of diisopropyl fluorophosphate (DFP) on the normal eye. *Arch. Ophthalmol.* **1946**, *36*, 17–32. [CrossRef]
45. Vorontsov, A.V.; Lion, C.; Savinov, E.N.; Smirniotis, P.G. Pathways of photocatalytic gas phase destruction of HD simulant 2-chloroethyl ethyl sulfide. *J. Catal.* **2003**, *220*, 414–423. [CrossRef]
46. Snijders, H.; Janssen, H.G.; Cramers, C. Optimization of temperature-programmed gas chromatographic separations I. Prediction of retention times and peak widths from retention indices. *J. Chromatogr. A* **1995**, *718*, 339–355. [CrossRef]
47. Shaban, M.T.; Baur, C.; Navab, N.; Albarqouni, S. Staingan: Stain style transfer for digital histological images. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (Isbi 2019), Venice, Italy, 8–11 April 2019; pp. 953–956.
48. Hauke, J.; Kossowski, T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* **2011**, *30*, 87–93. [CrossRef]
49. Kim, S.; Zhang, X. Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Comput. Math. Methods Med.* **2013**, *2013*, 509761. [CrossRef] [PubMed]
50. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
51. Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training generative adversarial networks with limited data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12104–12114.
52. Visa, S.; Ramsay, B.; Ralescu, A.L.; Van Der Knaap, E. Confusion matrix-based feature selection. *Maics* **2011**, *710*, 120–127.
53. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
54. Nettleton, D. Chapter 6—Selection of Variables and Factor Derivation. In *Commercial Data Mining*; Nettleton, D., Ed.; Morgan Kaufmann: Boston, MA, USA, 2014; pp. 79–104. [CrossRef]
55. LibreTexts, C. Quantitative and Qualitative GC and GC-MS. 2023. Available online: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Chromatography/Specific_Types_of_Chromatography/Gas_Chromatography/Quantitative_and_Qualitative_GC_and_GC-MS (accessed on 1 July 2024).
56. Technologies, A. Understanding Your Peaks: A Guide to Peak Detection and Integration in Chromatography. 2006. Available online: https://www.agilent.com/cs/library/technicaloverviews/public/5989-3425EN.pdf (accessed on 1 July 2024).
57. Jwaili, M. Pharmaceutical applications of gas chromatography. *Open J. Appl. Sci.* **2019**, *9*, 683–690. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.