

Review

Recent Advances in Generative Adversarial Networks for Gene Expression Data: A Comprehensive Review

Minhyeok Lee 

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea; mlee@cau.ac.kr

Abstract: The evolving field of generative artificial intelligence (GenAI), particularly generative deep learning, is revolutionizing a host of scientific and technological sectors. One of the pivotal innovations within this domain is the emergence of generative adversarial networks (GANs). These unique models have shown remarkable capabilities in crafting synthetic data, closely emulating real-world distributions. Notably, their application to gene expression data systems is a fascinating and rapidly growing focus area. Restrictions related to ethical and logistical issues often limit the size, diversity, and data-gathering speed of gene expression data. Herein lies the potential of GANs, as they are capable of producing synthetic gene expression data, offering a potential solution to these limitations. This review provides a thorough analysis of the most recent advancements at this innovative crossroads of GANs and gene expression data, specifically during the period from 2019 to 2023. In the context of the fast-paced progress in deep learning technologies, accurate and inclusive reviews of current practices are critical to guiding subsequent research efforts, sharing knowledge, and catalyzing continual growth in the discipline. This review, through highlighting recent studies and seminal works, serves as a key resource for academics and professionals alike, aiding their journey through the compelling confluence of GANs and gene expression data systems.

Keywords: generative adversarial networks (GAN); gene expression data; transcriptome; RNA; mRNA; deep learning; genomic data; artificial intelligence; generative artificial intelligence; genetics

MSC: 68T05



Citation: Lee, M. Recent Advances in Generative Adversarial Networks for Gene Expression Data: A Comprehensive Review. *Mathematics* **2023**, *11*, 3055. <https://doi.org/10.3390/math11143055>

Academic Editors: Sanghyuk Lee and Haolan Zhang

Received: 26 May 2023

Revised: 7 July 2023

Accepted: 10 July 2023

Published: 10 July 2023



Copyright: © 2023 by the author. 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

We are witnessing an epoch of generative artificial intelligence (GenAI), where the boundaries of creativity are being profoundly redefined. The emergence of generative deep learning models, notably the groundbreaking Generative Pretrained Transformers (GPT) [1–4] and diffusion models [5–9], has ushered in a new frontier where the very boundaries of creative expression are being reimaged. These cutting-edge advancements boldly challenge the longstanding belief that creativity is solely reserved for human intellect, unraveling a plethora of uncharted possibilities that lay before us, waiting to be explored.

In the vast array of deep learning-based GenAI, generative adversarial networks (GANs) have carved out a unique and noteworthy place [10–18]. GANs, conceptualized a decade ago [19], stand out as one of the most influential paradigms in deep-learning-based generative models. The versatility and robustness of GANs have engendered a multitude of applications across various scientific and technological fields. The defining feature of GANs, which sets them apart from other machine learning models, is their ability to generate synthetic data that closely approximates real data distributions, thus enriching the quality and diversity of available data.

A GAN is essentially a two-player game composed of a generator and a discriminator [20–22]. The generator's role is to create synthetic data, while the discriminator's task is to distinguish between real and generated data. During the training process, the

generator strives to produce data that the discriminator cannot differentiate from the real data, whereas the discriminator continually improves its ability to distinguish real from generated data. This adversarial training regimen imbues GANs with the capability to model complex data distributions and produce high-quality synthetic data.

Gene expression data are fundamentally constrained by the limitations imposed by the ethical and logistical challenges associated with human subject research [23]. In stark contrast to other domains that have made substantial progress harnessing the power of big data, gene expression data are limited in size, variety, and the rate at which they can be collected. However, GANs promise a potential solution to these obstacles by enabling the synthesis of a virtually limitless supply of artificial gene expression data.

This review aims to systematically explore and synthesize the recent advancements in the application of GANs for gene expression data. GANs, serving as a powerful data augmentation tool, allow for the artificial generation of copious amounts of gene expression data. This ability broadens the scope of exploration and application, igniting potential advancements in a myriad of applications within the domain of genomics and gene expression studies.

Our review focuses on the remarkable developments in this domain over the last five years, spanning from 2019 to 2023. Given the rapidly evolving nature of deep learning technologies, with advancements often unfolding on a monthly basis, it is of paramount importance to provide timely and comprehensive reviews. By highlighting the latest papers and groundbreaking research in the field, this review serves a crucial role in disseminating knowledge, stimulating further research, and contributing to the overall advancement of the field.

2. Literature Review

2.1. Paper Selection Methodology

The primary objective of the article selection process for this review was to ensure the inclusion of significant, high-quality research endeavors pertinent to the specific topic under investigation. To meet this goal, we utilized a robust algorithmic approach, primarily leveraging academic search platforms, with a special emphasis on the Web of Science (WOS). The devised search strategy involved meticulous selection of keywords such as “GANs”, “transcriptome”, “RNA”, and “gene expression”, tailored to capture relevant articles for this review.

Interestingly, while Variational Autoencoder (VAE) represents another significant GenAI model, our selection deliberately avoided papers focusing on VAE. The result of this rigorous process was an exhaustive list of 99 scholarly works published between 2019 and 2023. Nevertheless, a substantial portion of these publications did not precisely fall within the specific thematic scope we sought. Many studies focused more broadly on genetics rather than zeroing in on gene expression data. Further, a significant share of the reviewed articles strayed from the intended application of GANs.

In framing the scope of our review, we restricted our selection to articles published in peer-reviewed journals. This choice was driven by two primary considerations. Firstly, the peer review process lends credibility to scientific publications by subjecting them to the expert scrutiny of field specialists. Secondly, peer-reviewed journals are globally recognized as reputable sources for publishing scientifically rigorous and impactful research. Though we acknowledge the role of preprints and conference papers in academic discourse, our focus was specifically on peer-reviewed journal articles. This approach not only reinforces the soundness of our review, but also guarantees the inclusion of papers that have undergone a thorough review process.

In order to maintain the novelty and contemporariness of our review, we deliberately excluded article types such as review articles and perspectives, emphasizing the incorporation of original research-based studies. The selected articles span from 2019 to 2023, ensuring that the review reflects the most up-to-date trends and developments in the application of GANs to gene expression data.

We collected data until May 2023, adhering to our current timeline and ensuring that the review is up-to-date with recent advancements. During this phase, we collated information concerning the citation count and publication history for each article. These aspects were crucial for gauging the research’s scope, influence, and acceptance within the scientific community. To provide an organized overview of GANs’ application in gene expression data, we grouped the chosen articles based on their specific research objectives. This systematic classification fosters a comprehensive understanding of the genomic data deep learning landscape, enhancing the understanding of the diverse methodologies applied in this area. A summary of the articles reviewed is outlined in Table 1.

Table 1. Summary of recent investigations employing generative adversarial networks in gene expression data.

Research Topics	Brief Description	Studies
Gene Expression Data Generation	Techniques utilizing GANs to generate synthetic gene expression data, useful for boosting the performance of data-limited analyses	Yu et al. [24], Yelmen et al. [25], Hazra et al. [26], Zrimec et al. [27], Ahmed et al. [28], Vinas et al. [29], Marouf et al. [30], Chaudhari et al. [31], Kwon et al. [32], Mendez-Lucio et al. [33], Chen et al. [34]
Cancer Diagnosis	Employing GANs for cancer diagnosis, particularly improving sensitivity and specificity by addressing data imbalances	Tamilmani et al. [35], Xiao et al. [36], Wang and Zhang [37], Bi et al. [38], argonski et al. [39], Park et al. [40], Yang and Li [41], Gutta et al. [42], Mansoor et al. [43]
Single-Cell RNA-Seq Analysis	Utilization of GANs to handle single-cell RNA-seq data, including dropout imputation, batch effect removal, and feature extraction	Bahrami et al. [44], Jeon et al. [45], Wang et al. [46], Xu et al. [47], Wang et al. [48], Lin et al. [49], Xu et al. [50], Zhu et al. [51], Ding and Regev [52], Wei et al. [53], Reiman et al. [54], Wang et al. [55]
Genome Analysis	Applications of GANs in genome analysis, from creating artificial human genomes to variant calling	Wang et al. [56], Yang et al. [57], Kim et al. [58], Jiang et al. [59], Lee et al. [60], Du et al. [61], Yan et al. [62], Wang et al. [63]
Hi-C Data Enhancement	Using GANs to infer high-resolution Hi-C matrices from low-resolution counterparts, allowing detection of fine-scale chromatin interactions	Hu et al. [64], Hong et al. [65], Liu et al. [66], Liu et al. [67]
Other Applications	Various other applications of GANs in bioinformatics, including predicting modification sites and handling missing values in microarray data	Booker et al. [68], Salekin et al. [69], Pati et al. [70], Yuan et al. [71], Lee et al. [72], Kim et al. [73], Tsourtis et al. [74], Uthamacumaran [75], Xue et al. [76], Qiao et al. [77], Han et al. [78], Park et al. [79], Vinas et al. [80], Sui et al. [81]

2.2. Journals of Publications

The breadth and depth of work pertaining to GANs for gene expression are evident in the variety of journals that have published research on this topic between 2019 and 2023. These journals span several disciplines, highlighting the interdisciplinary nature of this field of study. Notably, the research articles reviewed here were published in a diverse set of thirty-four distinct journals. Figure 1 serves as a discerning visual representation, encapsulating the distribution of publications across various esteemed journals.

The most prolific journal, with a 14% share of the publications under review, is *Bioinformatics*. This underscores the journal’s role as a leading platform for disseminating original research in computational biology and bioinformatics, particularly in the arena of GANs applied to gene expression data.

Seven percent of the selected publications were made in *Nature Communications*, *PLoS Computational Biology*, and *IEEE Access*, respectively. Each of these journals has a distinct focus: *Nature Communications* emphasizes high-quality research from all areas of the natural sciences, *PLoS Computational Biology* features works of exceptional significance that further our understanding of living systems through the application of computational methods,

and *IEEE Access* is an interdisciplinary, open-access journal that publishes articles in the broad spectrum of engineering and technology.

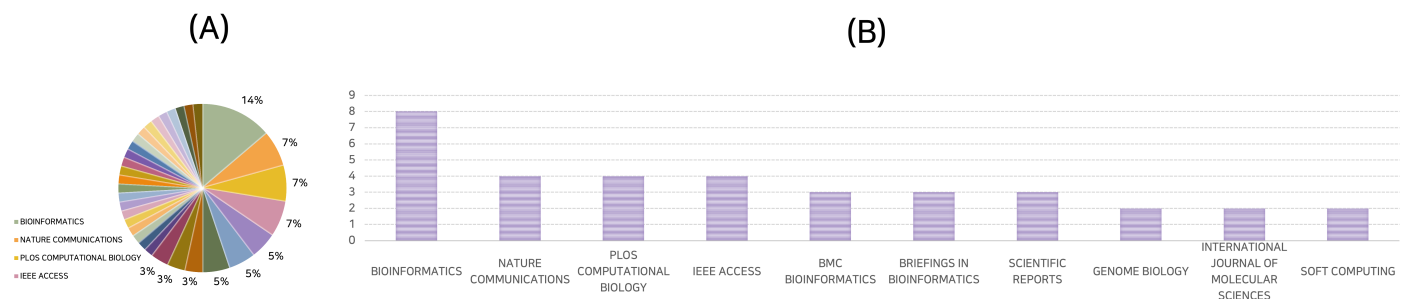


Figure 1. Representation of research papers across diverse journals. (A) Proportional illustration of papers in different journals; (B) Journals exhibiting multiple publications.

Journals such as *BMC Bioinformatics*, *Briefings in Bioinformatics*, and *Scientific Reports*, each with a 5% share of the reviewed publications, further emphasize the role of bioinformatics and interdisciplinary science in the investigation of GANs for gene expression data.

This analysis of the published journals provides an understanding of the academic venues that have been most receptive to research on GANs for gene expression. The wide range of journals reflects the interdisciplinary nature of this research area, which sits at the intersection of bioinformatics, genomics, machine learning, and artificial intelligence. The distribution of publications across these journals, as shown in Table 2, testifies to the importance and widespread recognition of this innovative research field.

Table 2. Journal statistics.

Journal Name	Num. of Papers	Percentage (%)
<i>Bioinformatics</i>	8	14%
<i>Nature Communications</i>	4	7%
<i>PLoS Computational Biology</i>	4	7%
<i>IEEE Access</i>	4	7%
<i>BMC Bioinformatics</i>	3	5%
<i>Briefings in Bioinformatics</i>	3	5%
<i>Scientific Reports</i>	3	5%
<i>Genome Biology</i>	2	3%
<i>International Journal of Molecular Sciences</i>	2	3%
<i>Soft Computing</i>	2	3%
Others (<2 Papers)	23	40%

2.3. Numerical Analysis of Selected Papers

To capture the momentum and growth trajectory of research in the field of GANs for gene expression data, a year-by-year breakdown of the selected papers was conducted. This exercise provided valuable insights into the research dynamics in this domain over the past five years.

As illustrated in Table 3, the field has experienced a marked growth in publications. In 2019, only a single paper was identified that fulfilled the selection criteria. This modest beginning was followed by a significant increase in research output in the subsequent years. The number of relevant publications jumped to 13 in 2020, reflecting a growing interest and recognition of the potential of GANs in gene expression studies. A comprehensive examination of the citation frequency and publication years for each paper is presented in Figure 2, providing a detailed analysis of these key metrics.

In 2021, the momentum was sustained, with the number of publications increasing to 15. The year 2022 represented a pivotal point, witnessing a surge in research output with 24 published papers. This demonstrates an impressive almost 60% increase from the

previous year and underscores the rapid acceleration in research interest and activities in this domain.

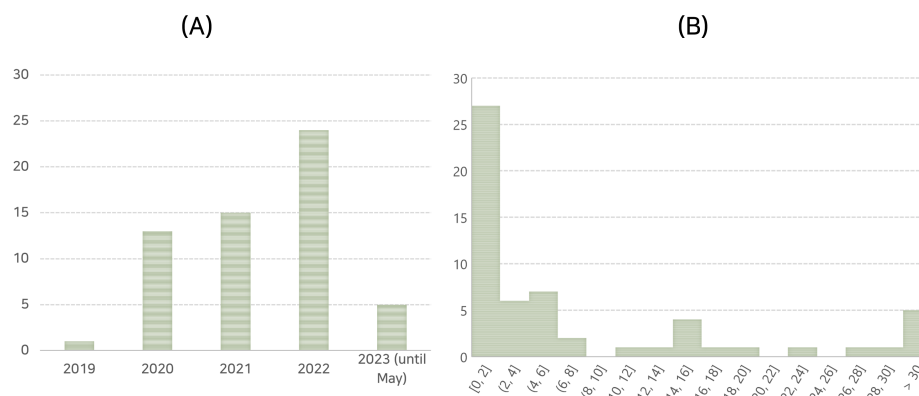


Figure 2. Overview of the distribution of publication years and frequencies of citations. (A) Distribution of publication years. (B) Distribution of citations frequencies.

Table 3. Yearly distribution of selected publications.

Year	Number of Publications
2019	1
2020	13
2021	15
2022	24
2023 (until May)	5

Even though data for 2023 is limited to publications until May, five papers have already been identified, signaling an ongoing upward trend. The dynamic growth pattern of this field, as evidenced by the increasing number of annual publications, emphasizes the rising importance of GANs for gene expression data in the research community.

In terms of the impact of these papers, the citation analysis revealed a median of three citations per paper and a mean of 9.7. This discrepancy between the median and mean suggests a right-skewed distribution of citations, indicating the existence of a few highly-cited papers. These influential papers have significantly contributed to shaping the research discourse in this field and have been instrumental in its rapid evolution.

3. Application of Generative Adversarial Networks in Gene Expression Data Augmentation

In the intersection of bioinformatics and artificial intelligence, GANs have emerged as innovative tools for the generation of gene expression data [82]. The GAN paradigm, equipped with the capacity to create synthetic data that mirrors real-world data, offers an appealing solution to inherent challenges faced in gene expression studies, including high dimensionality, sparse data, and sample diversity.

Research within this scope primarily aims to expand the breadth of existing gene expression datasets through the creation of synthetic data, offering an advanced alternative to conventional data augmentation or sampling methods. In contrast to these traditional methods that may distort the inherent data distribution, GANs excel in capturing and emulating the complex, non-linear patterns in gene expression datasets, leading to a more representative and reliable synthetic dataset. Such enriched datasets can subsequently enhance the efficacy of downstream analytical processes and predictive models. Besides expanding data quantity, the synthetic gene expression data generated by GANs can contribute significantly towards understanding and controlling data quality. GANs, with their ability to create data similar to the real-world distributions, can provide nuanced insights into the underlying biological phenomena. This knowledge can contribute to diverse areas,

including disease diagnosis, pharmaceutical development, and toxicogenomics, among others. The diagram depicted in Figure 3 showcases the schematic depiction of the GAN-based framework employed for data augmentation. Table 4 offers a comprehensive compilation of recent investigations conducted within the field.

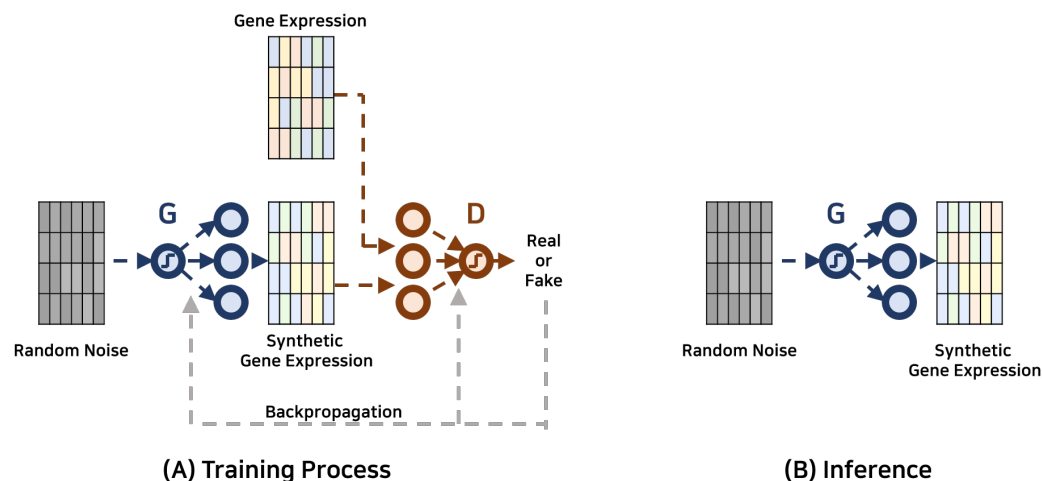


Figure 3. Gene expression data augmentation with generative adversarial networks. The generator and discriminator are represented by G and D, respectively.

Table 4. Contributions of different studies in the application of GANs in genomic data analysis.

Author	Contributions
Yu et al. [24]	Developed MichiGAN, a network combining VAEs and GANs to generate disentangled single-cell gene expression data.
Yelmen et al. [25]	Utilized GANs and RBMs to generate artificial human genomes enhancing data imputation quality for low frequency alleles.
Hazra et al. [26]	Used GANs to create synthetic nucleic acid sequences of the cat genome, achieving high correlation with original data.
Zrimec et al. [27]	Prototyped ExpressionGAN, generating synthetic regulatory DNA with targeted mRNA levels, exceeding natural controls in expression.
Ahmed et al. [28]	Introduced omicsGAN, integrating multi-omics data, enhancing predictive signals for cancer outcomes in synthetic data.
Vinas et al. [29]	Utilized a conditional GAN for generating realistic transcriptomics data preserving tissue- and cancer-specific properties.
Marouf et al. [30]	Developed cscGAN, generating realistic single-cell RNA-seq data, enhancing marker gene detection and classifier reliability.
Chaudhari et al. [31]	Developed MG-GAN to augment gene expression data, enhancing cancer classification accuracy significantly.
Kwon et al. [32]	Used GAN for augmenting samples, enhancing prediction of cancer stages significantly.
Mendez-Lucio et al. [33]	Introduced GAN model generating molecules inducing desired transcriptomic profile, a promising approach to drug discovery.
Chen et al. [34]	Developed Tox-GAN, generating gene activities and expression profiles, aiding in chemical-based read-across.

3.1. Recent Studies: 2019–2023

In the domain of manipulation and enhancement of gene expression data, several studies stand out. Yu et al. [24] designed the MichiGAN model by synergistically integrating Variational Autoencoders (VAEs) and GANs. The unique advantage offered by VAEs is their ability to create a latent space of variables, which captures the underlying

data distribution. This facilitates the generation of new samples by perturbing these latent variables, rendering VAEs apt for tasks requiring data augmentation. On the other hand, GANs, with their two-player adversarial framework, exhibit exceptional capabilities in generating high-quality, realistic data. In the MichiGAN model, the VAE component disentangles the representations of gene expression data, effectively mitigating the high dimensionality. The GAN component then works with these disentangled representations, generating synthetic samples that resemble real data. This seamless integration empowers the MichiGAN to generate high-quality synthetic gene expression data, thereby aiding in the prediction of cellular responses to drug treatments.

The strategy by Ahmed et al. [28] in developing omicsGAN lies in the integration of multi-omics data, specifically mRNA and microRNA expression data, and their interaction network. mRNA and microRNA are the two key components of gene regulation processes, providing a comprehensive view of the cell's functional elements. Integrating these data types, therefore, offers a broader and more accurate perspective of cellular states. OmicsGAN leverages this integrated data to generate synthetic data with enhanced predictive signals. The success of omicsGAN can be attributed to this integrative view of gene expression, which enables a more holistic capture of cellular phenomena, resulting in synthetic data that exhibits superior performance in predicting cancer outcomes.

Marouf et al. [30] showcased the application of conditional GANs in the creation of their cscGAN. The strength of conditional GANs resides in their ability to guide the data generation process by conditioning on auxiliary information. This allows the model to generate data with specific attributes, contributing to the generation of more targeted and meaningful data. In the case of cscGAN, the auxiliary information was single-cell RNA-seq data, which captures gene expression at an individual cell level, providing a more nuanced understanding of cellular heterogeneity. By conditioning the GAN on these data, cscGAN was able to generate realistic synthetic gene expression data at a single-cell resolution, enhancing downstream analyses and classifier reliability.

Both Yelmen et al. [25] and Hazra et al. [26] utilized GANs for the generation of synthetic genetic data, albeit with different source data. In the case of Yelmen et al. [25], artificial human genomes were generated, while Hazra et al. [26] synthesized nucleic acid sequences of the cat genome. The performance of these models is heavily reliant on GANs' innate ability to capture and reproduce the complex patterns inherent in genetic data. GANs, with their generator–discriminator structure, learn the data's intricate distributions, and are thereby capable of synthesizing high-quality genetic sequences that closely mirror the authentic data. This proficiency in learning from and replicating the real-world data distributions is key to their successful application in synthetic genetic data generation.

The central theme in the works of Chaudhari et al. [31] and Kwon et al. [32] is the use of GANs for gene expression data augmentation to enhance cancer classification. The choice of GANs for data augmentation stems from their ability to produce novel synthetic data that not only enlarges the dataset, but also reflects the true data distribution. The Modified Generator GAN (MG-GAN) by Chaudhari et al. is particularly effective because it uses a Gaussian distribution in the generator, thereby promoting a better emulation of the real data distribution. This results in higher-quality synthetic data, leading to an improvement in the accuracy of cancer classification.

In the context of drug discovery and toxicology, the success of the models by Mendez-Lucio et al. [33] and Chen et al. [34] can be attributed to the effective application of GANs to gene activities and expression profiles. GANs are proficient at learning from complex multi-dimensional data distributions and reproducing them. By applying GANs to gene activities and expression profiles, these models manage to generate synthetic data that accurately represents the biological phenomena.

3.2. Trends, Challenges, and Future Directions in Recent Studies

Several compelling trends have emerged from recent studies exploring the application of GANs in gene expression data. A prevalent theme revolves around the manipulation

and enhancement of gene expression data to improve its predictive power, as illustrated in the studies by Yu et al. [24] and Ahmed et al. [28]. These studies highlight the increasing trend toward harnessing the strengths of GANs, often in combination with other techniques such as VAEs, to generate synthetic data with enhanced predictive signals for biological and clinical outcomes.

The generation of synthetic genetic data, which closely mimics authentic genomic datasets, is another prominent trend. Work by Yelmen et al. [25] and Hazra et al. [26] exemplifies this direction, leveraging GANs and other deep learning architectures to generate realistic, artificial human and cat genomes, respectively. These synthetic datasets have proven instrumental in enhancing data imputation quality and fostering an improved understanding of complex genomic structures and functions.

Despite these advancements, a number of challenges persist. One primary concern is the need to develop more rigorous measures for evaluating the performance and reliability of GANs in gene expression analysis. Currently, there is a lack of standard evaluation metrics and validation datasets that can provide objective assessments of these models. In addition, while GANs are incredibly powerful, their complexity can make them difficult to interpret and apply in practice. Exploring ways to make these models more interpretable and user-friendly will be crucial to their broader adoption in the biological and clinical community.

Looking ahead, there are numerous intriguing prospects for future research. The application of GANs in drug discovery, as illustrated by Mendez-Lucio et al. [33], provides a novel avenue for developing more effective therapeutics. The same applies to the work of Chen et al. [34], where GANs were utilized to generate gene activities and expression profiles, demonstrating their potential utility in toxicogenomics. Therefore, one promising future direction may lie in further expanding the scope of GANs in translational genomics, potentially revolutionizing drug discovery, therapeutic strategies, and personalized medicine. Moreover, combining GANs with other machine learning and deep learning techniques could foster even more powerful and versatile tools for genomic data analysis. As these trends evolve, it will be exciting to see how GANs continue to reshape the landscape of genomics and bioinformatics.

4. Generative Adversarial Networks for a Novel Approach to Cancer Diagnosis

In the domain of oncological research and diagnosis [83], GANs are demonstrating their potential as a game-changing tool. Harnessing the capability of GANs to learn intricate data distributions and generate synthetic data, researchers are finding novel ways to analyze cancer-related gene expression data, thereby paving the path towards improved detection and classification of various cancer types.

Cancer diagnosis frequently involves the analysis of complex and high-dimensional data, such as gene expression profiles or medical imaging data. GANs, with their proficiency in discerning patterns in complex datasets, are being increasingly used to supplement the available data by generating synthetic data. This synthetic data can be used to train other machine learning models, enhancing their performance in cancer diagnosis.

In this context, research endeavors focus on developing GAN-based frameworks capable of interpreting gene expression data or medical images to diagnose different types of cancer. GANs' ability to model non-linear mappings allows them to capture the complex relationships within high-dimensional data. This aptitude for modeling data intricacies makes GANs particularly effective in distinguishing various cancer types based on gene expression patterns or imaging data.

In addition, GANs can unearth subtle cancer characteristics by generating synthetic data that mirrors real-world distributions, offering an improved understanding of the complex nature of cancer. These insights can contribute to the refinement of diagnosis and treatment strategies. Despite these advantages, the implementation of GANs in cancer diagnosis is not without its challenges, including concerns regarding the interpretability of GAN models, the biological validity of the synthetic data, and the risk of overfitting.

Therefore, constant advancements in methodologies, rigorous assessment of synthetic data, and strategies to improve model transparency are required in this promising but challenging field. The visual representation presented in Figure 4 serves as an exemplification of the conceptual framework utilized in the data imbalance problem for cancer diagnosis, which utilizes GANs. A succinct representation of contemporary research endeavors is portrayed in Table 5.

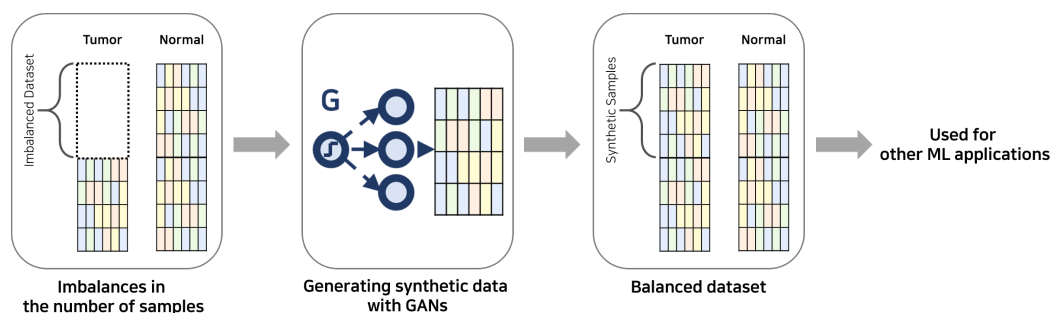


Figure 4. Generative adversarial networks in cancer diagnosis to address data imbalance. The generator is represented by G.

Table 5. Contributions of various studies leveraging generative adversarial networks for cancer diagnosis.

Author	Contributions
Tamilmani et al. [35]	Developed an optimized GAN for accurate cancer miRNA biomarker classification.
Xiao et al. [36]	Presented an improved deep learning based WGAN model for imbalanced gene expression data in cancer diagnosis.
Wang and Zhang [37]	Created a novel predictor for lymph node metastasis events in tumors using lncRNA expression profiles.
Bi et al. [38]	Proposed FIAD-GAN, a GAN framework for brain disease classification and feature selection.
Targonski et al. [39]	Introduced TSPG, a GAN-based method to identify transcriptome perturbations in RNA-sequencing data.
Park et al. [40]	Presented a method using graph-embedded GAN for accurate cancer prognosis prediction.
Yang and Li [41]	Introduced BiGAN for predicting lncRNA-disease associations using similarity measures.
Gutta et al. [42]	Utilized a WGAN with gradient penalty to enhance breast cancer prognostication.
Mansoor et al. [43]	Proposed GOGAN, a novel architecture for predicting protein functions using amino acid sequences.

4.1. Recent Studies: 2019–2023

In the case of Tamilmani et al. [35], the success of the proposed DCG-MCNN model is attributed to its method of handling superclass problems. In cancer diagnoses, the nature of data often results in a high number of superclasses, i.e., groups of samples that are difficult to distinguish based on the available features. DCG-MCNN employs a GAN optimized using the Mayfly Optimization Algorithm, a nature-inspired optimization technique that utilizes the “group effect” phenomenon observed in mayflies to solve complex problems. By using this optimization technique, the authors were able to train the GAN to more effectively generate synthetic data, which in turn improved the classification of superclasses. The application of a deep convolutional generative adversarial network (DCG) and a modified convolutional neural network (MCNN) further enhanced this process. DCG is a type of GAN that uses convolutional layers in its generator and discriminator networks, allowing for better feature extraction from high-dimensional data, such as gene expression profiles. MCNN, on the other hand, involves modifications in the architecture of traditional convolutional neural networks to better manage the complexity of data and improve classification accuracy. The interplay between these components in the DCG-MCNN model resulted in the model’s superior performance.

In contrast, Xiao et al. [36] addressed the challenge of imbalanced gene expression data, a common problem in cancer diagnosis, using a Wasserstein GAN model (WGAN). The WGAN is known for its theoretical improvements over the original GAN model, particularly in terms of stability during training. This stability is achieved by replacing the traditional GAN loss function with the Wasserstein distance, a more meaningful measure of the difference between the true data distribution and the distribution learned by the GAN. The authors used this unique characteristic of the WGAN to generate new samples for the minority class, thereby addressing the data imbalance. By enriching the minority class in this way, the authors were able to achieve more balanced data distributions, which in turn led to improved prediction accuracy.

The study of Wang and Zhang [37] also made use of the WGAN, but coupled it with a particle swarm optimization neural network (psoNN) for predicting metastasis in tumors. The psoNN is an algorithm that uses the concept of swarm intelligence, where a group of simple agents collectively perform complex tasks, to optimize the parameters of a neural network. By integrating this optimization method with the WGAN, the authors were able to train a model that effectively captured the complex relationships within lncRNA expression profiles. This innovative approach to data modeling allowed the WGAN-psoNN to predict metastasis events with high accuracy.

In the study conducted by Bi et al. [38], the FIAD-GAN framework proposed was able to efficiently aggregate feature information and perform convolution/deconvolution operations. This dual approach provided the model with the ability to extract relevant information from the gene expression data, thus improving the interpretability of the deep learning framework. The enhanced interpretability, in turn, provided a more in-depth understanding of the disease-related brain regions and genes, thus contributing to the model's superior performance in disease classification tasks.

The success of the TSPG method presented by Targonski et al. [39] can be attributed to the way it leverages GANs for transcriptome perturbation analysis. The TSPG method uses GANs to identify the necessary transcriptome perturbations to transition between classes, providing a deeper understanding of the differential expression patterns in normal and tumor human tissue samples. This nuanced understanding is essential in precision medicine applications, where identifying substantial transcriptional aberrations can provide valuable insights into the development and progression of diseases.

The graph-embedded GAN proposed by Park et al. [40] successfully captured the distributions of various data types, including gene expression, copy number variation, single nucleotide polymorphism, and DNA methylation data. The integration of a graph-embedded GAN in this study allowed for the extraction of relational structures in the data, a feature that is particularly useful when dealing with genetic data where genes often interact in complex ways. By learning these distributions, the graph-embedded GAN provided an improved representation of the data, leading to higher prediction accuracy.

The BiGAN model presented by Yang and Li [41] achieves its performance by utilizing disease semantic similarity, lncRNA sequence similarity, and Gaussian interaction profile kernel similarities to construct feature representations. This three-pronged approach captures multiple aspects of lncRNA-disease associations, providing a comprehensive understanding of these relationships. This multi-faceted understanding enables BiGAN to accurately predict associations, outperforming other methods in cross-validation and case studies.

Gutta et al. [42] utilized a Wasserstein GAN with a gradient penalty (T-GAN-D) to improve breast cancer prognostication. The gradient penalty enforced the Lipschitz constraint on the discriminator, ensuring its 1-Lipschitz continuity. This continuity is crucial in stabilizing the GAN training process, thereby enhancing the quality of the generated synthetic data. This high-quality synthetic data, when applied to a large cohort of breast cancer patients, resulted in superior differentiation between high-risk and low-risk patients compared to traditional biomarkers.

The GOGAN model proposed by Mansoor et al. [43] works by extracting features from large unlabeled protein datasets. The extraction of these features provides valuable insights into the properties of the proteins, which can then be used to predict protein functions. The GOGAN model's ability to learn these features and incorporate them into the prediction process results in its superior performance in terms of various evaluation metrics. The model's ability to analyze gene variations, gene expression, and gene regulation networks further emphasizes the effectiveness of this feature extraction method in predicting protein functions.

4.2. Trends, Challenges, and Future Directions in Recent Studies

The preceding analyses offer a panorama of recent studies leveraging GANs in cancer diagnosis and related genomics research, reflecting both considerable advancements and persistent challenges in the field. A visible trend is the innovative use of GANs to enhance the quality and quantity of genomic data, as evidenced in Tamilmani et al.'s [35] DCG-MCNN and Xiao et al.'s [36] WGAN model, both of which address data imbalance issues. On another frontier, studies like Targonski et al.'s [39] explore the utility of GANs in identifying transcriptome perturbations, broadening the horizon of GAN application in genomic data analysis.

Yet, despite these advancements, the field faces several challenges. For one, the complexity and high dimensionality of genomic data pose difficulties for the successful training of GANs. Overfitting and mode collapse, common issues in GAN models, can be exacerbated in genomics contexts due to the intricate structure of the data. The issue of data imbalance is another significant challenge, as reflected in the works of Xiao et al. [36] and Wang and Zhang [37], warranting further methodological advancements in GAN architectures to address this concern.

Future research should aim to resolve these challenges and continue to harness the potential of GANs in genomics. Combining GANs with other machine learning techniques, as demonstrated by Tamilmani et al.'s [35] DCG-MCNN and Bi et al.'s [38] FIAD-GAN, is a promising direction. This line of inquiry could lead to the development of more efficient hybrid models that enhance the robustness and reliability of genomics research. Additionally, efforts to improve the interpretability and transparency of GAN models, like those seen in Bi et al.'s [38] work, are necessary to facilitate the broader adoption of these methods. As GANs continue to evolve, their potential to advance our understanding of complex biological systems becomes ever more promising.

5. Exploiting Generative Adversarial Networks in Single-Cell RNA-Seq Data Analysis

GANs have made significant strides in the sphere of single-cell RNA sequencing (scRNA-Seq) analysis [84–86], showcasing their ability to navigate the complexities of high-dimensional biological data. scRNA-Seq, an innovative genomics technology, provides granular insights into cell-specific gene expression. However, scRNA-Seq data's high dimensionality, sparsity, and inherent noise pose substantial obstacles in data interpretation and analysis. In this regard, GANs have been identified as a potent analytical tool.

The core application of GANs in scRNA-Seq analysis involves the generation of synthetic data, which can ameliorate issues related to data scarcity and noise that are typically associated with scRNA-Seq studies. The generated synthetic data, while reflecting the distribution of the actual scRNA-Seq data, can increase the reliability of downstream analysis and fortify the effectiveness of subsequent predictive models.

Research initiatives in this domain focus on the development and application of GAN architectures that are capable of modeling the distribution of scRNA-Seq data effectively. The primary aim is to capture the variability and biological complexity inherent in the data, thus creating synthetic cells that mirror the real ones. This amplification of available data for analysis can foster a deeper understanding of cell heterogeneity and cellular functions.

An additional avenue worth exploring is the application of Generative Adversarial Imputation Networks (GAIN) in the imputation of scRNA-Seq data. GAIN is an innovative

technique, marrying the strengths of GANs and imputation methodologies, which can be extremely useful in dealing with data sparsity, a common problem with scRNA-Seq data [87]. It operates by modeling the data distribution, much like traditional GANs, and employs a generator and a discriminator network to predict and evaluate the missing values, respectively. The generator, trained to produce realistic imputations for missing data points, minimizes the difference between the original and imputed data, thereby preserving the distribution of the original scRNA-Seq data. The discriminator, on the other hand, is trained to distinguish between observed and imputed data, resulting in a robust imputation model that can effectively navigate the inherent challenges posed by the high-dimensionality of scRNA-Seq data. Consequently, GAIN can significantly enhance the quality of data imputation, and in doing so, augment the reliability of subsequent data analysis, which is critical in advancing our understanding of complex biological phenomena.

Beyond mere data augmentation, GANs in scRNA-Seq analysis also promise potential breakthroughs in other areas. For instance, their ability to generate synthetic data, accurately reflective of real-world distributions, could reveal subtle cellular states and transitions, typically obscured in conventional analysis. This capability can potentially pave the way for new biological insights and advancements in areas such as developmental biology and disease research. Figure 5 provides a graphical representation that showcases the framework utilized in gene expression data imputation, leveraging GANs. Table 6 presents a concise snapshot of recent studies conducted in the field.

Table 6. Contributions of recent studies in single-cell RNA-Seq data analysis using generative adversarial networks.

Author	Contributions
Bahrami et al. [44]	Introduced scGAN to mitigate batch effects in scRNA-seq data, improving gene association identification and cell type clustering.
Jeon et al. [45]	Developed a two-step model that transforms L1000 profiles to RNA-seq-like profiles, enhancing the accuracy of gene profile translation.
Wang et al. [46]	Proposed iMAP, a framework that removes batch effects in single-cell RNA-sequencing datasets, uncovering novel cell–cell interactions in tumor environments.
Xu et al. [47]	Introduced scIGANs to tackle dropout imputation in scRNA-seq data, enhancing downstream analysis without oversmoothing.
Wang et al. [48]	Developed IMG2 to integrate multiple scRNA-seq datasets, enhancing downstream tasks like differential gene expression analysis.
Lin et al. [49]	Introduced DR-A for dimensionality reduction in scRNA-seq data, improving the accuracy of low-dimensional data representation.
Xu et al. [50]	Proposed scSemiGAN for single-cell RNA-seq data analysis, unifying cell-type annotation with dimensionality reduction and deep latent representation learning.
Zhu et al. [51]	Utilized scIGANs in imputation procedures to improve the analysis of single-cell RNA-seq data, enhancing clustering and differential expression analyses.
Ding and Regev [52]	Introduced scSphere to embed scRNA-seq profiles into low-dimensional spaces, enabling effective visualization and elucidation of complex cellular relationships.
Wei et al. [53]	Developed scPreGAN to predict the response of single-cell expression to perturbation, outperforming previous methods.
Reiman et al. [54]	Created Pseudocell Tracer to infer cellular trajectories from scRNAseq data, revealing key transcription factor ordering and poised cytokine signaling states.
Wang et al. [55]	Proposed ResPAN for batch correction in scRNA-seq data integration, outperforming other methods in both batch correction and biological information conservation.

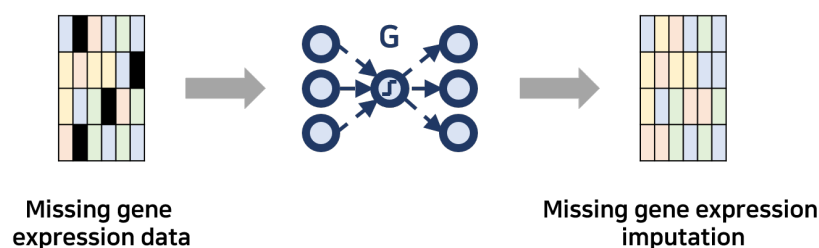


Figure 5. Generative adversarial networks for missing gene expression data imputation. The generator is represented by G.

5.1. Recent Studies: 2019–2023

Understanding the principles and mechanisms that underpin the success of GANs in scRNA-seq data analysis requires a comprehensive look at the challenges posed by the analysis of scRNA-seq data and how the architecture and functionality of GANs address these challenges. Here, we explore some of these principles by focusing on the key studies discussed above.

The scGAN, introduced by [44], proved effective in mitigating batch effects in scRNA-seq data due to its ability to accurately model the distribution of single-cell transcriptomes. Batch effects are systematic non-biological differences between batches (or groups) of samples in high-throughput experiments and can be extremely problematic in scRNA-seq analysis. ScGAN leverages the generator component of the GAN to produce synthetic data that follow the same underlying distribution as the real data, effectively removing batch-specific variations and retaining only the common features across batches. The discriminator, trained to distinguish between real and synthetic data, provides a feedback mechanism that continually refines the generator's outputs. This continual interaction between the generator and discriminator in the GAN structure allows for a nuanced understanding and accurate modeling of the complex data distributions in scRNA-seq datasets, thereby mitigating batch effects.

The success of iMAP [46] in uncovering novel cell–cell interactions lies in its innovative use of deep autoencoders and GANs. Deep autoencoders, as unsupervised learning algorithms, are trained to reconstruct the input data after reducing it to a lower-dimensional representation. They can therefore identify and capture the key features of the scRNA-seq data. Meanwhile, GANs complement this by generating synthetic data that mirror the data distributions of real cells. The combined effect is a comprehensive representation of cell-specific gene expressions, which allows for the discovery of novel interactions that would otherwise be obscured in traditional analyses.

Addressing dropout events in scRNA-seq data is another significant challenge, effectively handled by scIGANs as proposed by Xu et al. [47]. Dropouts, which refer to the missing values in scRNA-seq data due to low amounts of mRNA in a cell, can severely impact downstream analysis. GANs are uniquely suited to address this issue as they can model complex data distributions and generate synthetic data to fill in for the dropouts. The generator in scIGANs produces synthetic cells for dropout imputation, effectively mitigating the common issues of oversmoothing and stochasticity elimination in traditional imputation methods. The discriminator's role in distinguishing between observed and imputed data ensures that the imputed data points do not deviate significantly from the original data distribution, thus preserving the integrity of the scRNA-seq data.

In the context of integrating multiple scRNA-seq datasets, IMG [48] and ResPAN [55] have shown remarkable efficacy. These models use GANs to eliminate non-biological differences between batches while preserving biological information. Specifically, ResPAN employs the Wasserstein GAN, which uses a different loss function than traditional GANs, ensuring a more stable training process and resulting in more accurate generated data. This is crucial in maintaining the authenticity of the biological information while correcting for batch effects.

5.2. Trends, Challenges, and Future Directions in Recent Studies

An examination of recent studies reveals that GANs are being increasingly adopted for scRNA-seq data analysis. This trend represents an effort to overcome the specific challenges associated with scRNA-seq data, such as batch effects, dropout events, and high dimensionality, which have been hindrances in conventional analysis methods.

One major trend has been the application of GANs to mitigate batch effects, as demonstrated by Bahrami et al. [44] and Wang et al. [46,48,55]. These studies represent concerted efforts to improve the quality and consistency of scRNA-seq data across different experiments or conditions, enhancing the robustness and reliability of downstream analyses.

The challenge of dropout imputation has also seen significant attention, with innovative solutions such as scIGANs [47,51] being introduced. This development demonstrates a clear recognition of the limitations of traditional imputation methods, such as oversmoothing, which can undermine the granularity and biological accuracy of scRNA-seq data.

However, despite the substantial progress, several challenges persist. The complexity and variety of scRNA-seq data often require specific, tailored solutions. The proposed GAN models are often complex and require significant computational resources, which may limit their applicability in some cases. There is also a continuing need for methods that can handle larger, more diverse datasets, integrating information from multiple sources and types of data.

It is clear that the continued development and refinement of GAN-based models will play a significant role in the future of scRNA-seq data analysis. For instance, the works by Jeon et al. [45] and Lin et al. [49] indicate potential advancements in transforming gene expression profiles and enhancing dimensionality reduction, respectively. These trends suggest promising avenues for future research, with potential implications for our understanding of complex biological systems.

A future direction for research may also include the integration of GANs with other machine learning methods to create more sophisticated models. The model developed by Reiman et al. [54], for instance, shows how GANs can be used to generate pseudocells and infer cellular trajectories, providing a tantalizing glimpse of how these models could be further expanded and refined to uncover new biological insights.

While GANs have already made substantial contributions to the analysis of scRNA-seq data, they are far from reaching their full potential. The ongoing development of these models will continue to push the boundaries of our understanding of gene expression and cellular function, providing new insights into the complex interplay of genes and cells in various biological contexts.

6. Implementing Generative Adversarial Networks in Genome Analysis

The integration of GANs into genome analysis offers a novel strategy for understanding and interpreting genomic data. GANs' inherent competency in generating synthetic data, coupled with their capacity to model complex data distributions, makes them a formidable tool for analyzing genomic sequences. Generally, this application entails the utilization of GANs to generate synthetic genomic sequences, thus supplementing existing data sets and facilitating downstream analysis.

The use of GANs in genome analysis primarily aims to address the challenges posed by genomic data, which often involve high dimensionality, limited data availability, and intricate sequence patterns. GANs are used to create synthetic data that closely resembles the distribution of real genomic sequences. This synthetic data bolsters the robustness of genomic studies, providing a valuable resource for additional analysis and enhancing the performance of subsequent predictive models.

Research in this area is commonly centered around the design and implementation of GAN architectures capable of effectively modeling the distribution of genomic data. The principal goal is to capture the subtleties and complexities that are inherent in genomic sequences, thereby generating synthetic sequences that are representative of the actual

genomic landscape. The use of GANs thus provides a methodology for data augmentation, enriching the breadth and depth of genomic studies.

In addition to data augmentation, GANs offer potential advantages in other aspects of genome analysis. For example, GANs' ability to generate synthetic sequences that accurately reflect real-world genomic distributions may reveal nuanced genomic features and patterns. This potential can lead to deeper insights into the structure and function of genomes, thereby contributing to advancements in fields such as comparative genomics, population genetics, and functional genomics. The diagram provided in Figure 6 offers a visual representation of the GAN-based framework adopted in this academic inquiry. Table 7 furnishes an informative overview of contemporary research endeavors.

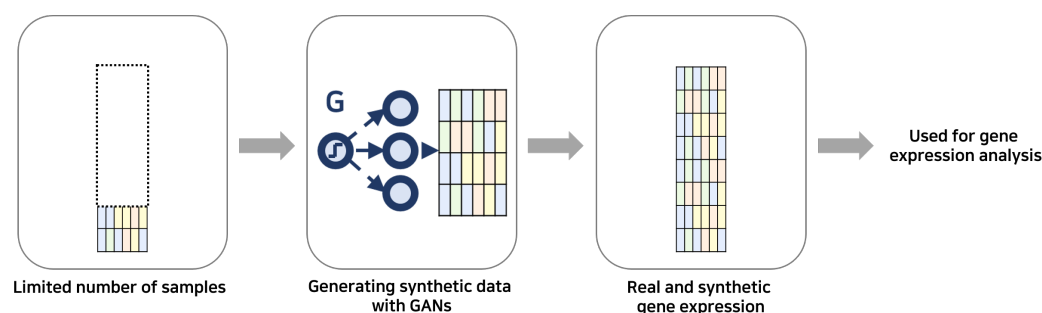


Figure 6. Framework of generative adversarial networks in genome analysis. The generator is represented by G.

Table 7. Summary of generative adversarial network contributions in genomics studies.

Author	Contributions
Wang et al. [56]	Developed pg-gan to accurately estimate parameters in population genetic models.
Yang et al. [57]	Developed LDV-Caller for genome variant calling from low-depth ONT sequencing data.
Kim et al. [58]	Utilized GANs to simulate molecular progression of tauopathy.
Jiang et al. [59]	Introduced a GAN model to predict disease genes from RNA-seq data.
Lee et al. [60]	Developed BP-GAN for branchpoint prediction in RNA sequences.
Du et al. [61]	Created LDA-GAN to predict lncRNA–disease associations.
Yan et al. [62]	Proposed GANCDAs, a model for predicting circRNA–disease associations.
Wang et al. [63]	Proposed SGANRDA, a model for predicting circRNA–disease associations, integrating circRNA sequence and disease semantics.

6.1. Recent Studies: 2019–2023

The pg-gan, proposed by Wang et al. [56], estimates parameters in population genetic models effectively because it leverages the ability of GANs to accurately model complex data distributions. In the context of population genetics, the real-world data often represent a high-dimensional and intricate distribution, which presents a challenge for traditional machine learning models. However, the generator component in pg-gan learns to replicate this complex distribution and generates synthetic data that closely resemble the actual genomic data. The discriminator, on the other hand, guides the generator's training by distinguishing between real and synthetic data, driving the generator to produce increasingly accurate synthetic data. This iterative process results in a more precise estimation of population genetic parameters.

The LDV-Caller [57] exhibits high accuracy in genome variant calling from low-depth ONT sequencing data, mainly due to the capacity of GANs to compensate for limited data availability. In low-depth sequencing, the scarcity of data can result in inaccurate variant calling. However, by training on real genomic variants, the generator in the LDV-Caller learns to produce synthetic variants that accurately reflect the distribution of real genomic

variants. Consequently, the synthetic data effectively supplement the sparse real data, leading to improved variant-calling accuracy.

In Kim et al. [58], GANs are used to simulate the molecular progression of tauopathy. The key to their success lies in the ability of GANs to uncover intricate patterns in the data. In the case of tauopathy, the early molecular changes can be subtle and easily overlooked by traditional analyses. However, GANs, by virtue of their capacity to model complex distributions and generate synthetic data, can replicate these early changes accurately. Furthermore, the continual interaction between the generator and discriminator enables the model to refine its understanding of the molecular progression, thereby unveiling early disease features effectively.

The efficacy of the GAN model proposed by Jiang et al. [59] in identifying disease genes from RNA-seq data can be attributed to its ability to handle small sample sizes. The generator in this model generates synthetic RNA-seq data that mimic the distribution of real disease genes. This artificially enlarges the dataset, mitigating the problems caused by small sample sizes. The discriminator, in turn, ensures that the synthetic data closely resemble the actual data, leading to improved accuracy in identifying disease genes.

The BP-GAN [60] stands out in predicting branchpoints in RNA sequences due to its ability to capture the latent structure of RNA sequences and the sequence-positional long-term dependency. The generator in BP-GAN learns to generate synthetic RNA sequences that reflect the intrinsic latent structure and long-term dependencies, while the discriminator ensures the quality of these synthetic sequences by distinguishing them from real sequences. This iterative process results in an increasingly accurate understanding of the RNA sequences, thereby leading to effective branchpoint prediction.

The predictive models for disease associations proposed by Du et al. [61], Yan et al. [62], and Wang et al. [63], namely LDA-GAN, GANCD, and SGANRDA, respectively, harness the strengths of GANs in data augmentation and complex distribution modeling. Limited data availability and complex data structures are common challenges in disease association prediction. However, GANs generate synthetic data that closely resemble the distribution of real associations, effectively overcoming the data limitation issue. Simultaneously, GANs' ability to model complex distributions ensures that the synthetic data accurately reflect the nuanced relationships between different genomic features and diseases, resulting in highly accurate prediction models.

6.2. Trends, Challenges, and Future Directions in Recent Studies

The recent developments in the application of GANs to genomics demonstrate a promising trend in tackling key challenges in the field. As the studies presented in Table 7 demonstrate, the utilization of GANs has significantly enhanced the accuracy and efficiency of tasks ranging from parameter estimation in population genetics to disease gene identification and association predictions.

The utilization of GANs in population genetics by Wang et al. [56] represents a key advancement in the field. However, it brings to light the inherent complexity and variability in genetic data. Future studies should focus on improving the versatility of GANs to accommodate different types of genetic data and populations. Additionally, there is a need for more robust methods for model evaluation and comparison to ensure the optimal application of GANs in this domain.

While the use of GANs for variant calling from low-depth sequencing data, as demonstrated by Yang et al. [57], marks an important advancement, the challenge of managing the vast amount of data generated through next-generation sequencing technologies persists. As such, future efforts should prioritize the development of more efficient data processing and storage techniques to complement the improved analytical capabilities provided by GANs.

In disease gene identification, GANs have proven instrumental in circumventing the challenge of small sample sizes in brain-related disease studies [59]. Nevertheless, more research is needed to tailor GAN models to other diseases with similar challenges.

Moreover, the issue of model interpretability remains critical. The black-box nature of GANs, while providing impressive predictive power, can limit our understanding of the mechanisms behind the predictions. Therefore, future research should focus on developing techniques to improve the interpretability and transparency of GAN models.

Finally, the application of GANs in disease association predictions, specifically in predicting lncRNA–disease and circRNA–disease associations [61–63], highlights the vast potential of GANs in predictive genomics. Yet, the limited availability of comprehensive and accurate disease association databases can restrict the effectiveness of GANs in this area. Therefore, more emphasis should be placed on the generation of high-quality databases to ensure reliable predictions.

7. Utilizing Generative Adversarial Networks for Hi-C Data Enhancement

GANs have found a noteworthy application in the domain of high-throughput chromosome conformation capture (Hi-C) data enhancement [88–90], showcasing their potential in tackling complex biological data. Hi-C is an innovative technology that offers insights into the three-dimensional (3D) organization of the genome within the nucleus. However, Hi-C data's high dimensionality, sparsity, and inherent noise present significant hurdles in its analysis. GANs are emerging as a powerful tool to mitigate these challenges.

The integration of GANs in Hi-C data enhancement primarily revolves around generating synthetic Hi-C data that resembles the distribution of real Hi-C matrices. These synthetic data can supplement existing Hi-C datasets, aiding in overcoming data scarcity issues, and improving the robustness of downstream analysis.

Research pursuits in this field primarily involve the development of GAN architectures that can effectively capture the complex data patterns inherent in Hi-C matrices. The overarching goal is to generate synthetic Hi-C data that retains the structural complexities of the actual genomic architecture. The augmented datasets can improve the efficacy of subsequent analyses and potentially enhance our understanding of genome folding patterns and their biological implications.

In addition to data augmentation, GANs can be instrumental in other facets of Hi-C data enhancement. For instance, synthetic data reflecting real-world 3D genomic distributions could uncover nuanced genomic architectural features, which may be overlooked in conventional Hi-C analyses. This potential could lead to a deeper understanding of 3D genome organization and its role in gene regulation and cellular functions. The diagram featured in Figure 7 visually elucidates the GAN-driven framework employed in Hi-C data enhancement. Table 8 exhibits a curated collection of recent studies undertaken within the field.

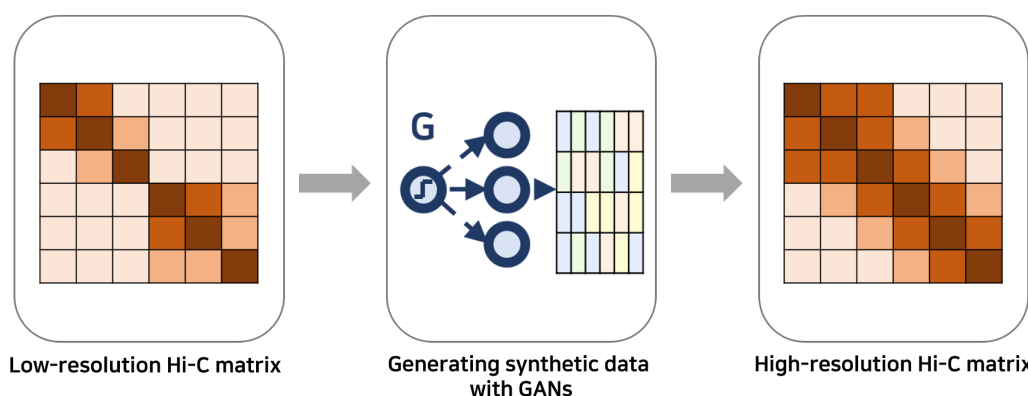


Figure 7. Generative adversarial networks for high-resolution Hi-C matrices. The generator is represented by G .

Table 8. Contributions of recent studies in generative adversarial networks for Hi-C data enhancement.

Author	Contributions
Hu et al. [64]	EnHiC: Enhancing the resolution of Hi-C matrices using a GAN framework, enabling the detection of topologically associated domains and fine-scale chromatin interactions.
Hong et al. [65]	DeepHiC: Predicting high-resolution Hi-C contact maps from low-coverage sequencing data using a GAN, outperforming other methods in chromatin loops identification and TADs detection.
Liu et al. [66]	hicGAN: Inferring high-resolution Hi-C data from low-resolution counterparts using GANs, producing matrices consistent with original high-resolution data and enhancing resolution in new cell types.
Liu et al. [67]	scDEC-Hi-C: Analyzing single-cell Hi-C data with a GAN-based framework, uncovering chromatin architecture variability across cell types with superior clustering and imputation performance.

7.1. Recent Studies: 2019–2023

The success of EnHiC [64] in predicting high-resolution Hi-C matrices from low-resolution data can be attributed to two major factors: the combination of GANs with non-negative matrix factorization (NMF) and the iterative feedback mechanism inherent to GANs. The former allows the model to extract key structural features from multi-scale low-resolution matrices, which are then used by the GAN's generator to produce high-resolution Hi-C matrices. The discriminator component of the GAN then evaluates the synthetic matrices against real high-resolution data, thus iteratively guiding the generator to produce more accurate predictions. This combination of feature extraction via NMF and synthetic data generation via GANs results in enhanced resolution of Hi-C matrices.

DeepHiC, developed by Hong et al. [65], achieves superior performance in predicting high-resolution Hi-C contact maps from low-coverage sequencing data primarily due to the GAN's ability to learn from limited and sparse data. The scarcity of data inherent in low-coverage sequencing can result in inaccurate predictions. However, the GAN's generator learns to produce synthetic contact maps that reflect the distribution of actual high-resolution data, effectively overcoming the data limitation. Simultaneously, the discriminator guides the generator's learning, ensuring the quality of the synthetic contact maps, which leads to the accurate detection of chromatin loops and topologically associated domains (TADs).

The hicGAN [66] is able to infer high-resolution Hi-C data from low-resolution counterparts primarily due to the model's ability to capture the intricate patterns inherent in Hi-C data. In the case of hicGAN, the generator learns to produce synthetic high-resolution Hi-C matrices that mimic the distribution of actual high-resolution data. The discriminator, in turn, guides the generator's training, distinguishing between real and synthetic data, driving the generator to produce increasingly accurate high-resolution matrices. This iterative feedback process results in the generation of high-resolution Hi-C data that is consistent with original high-resolution matrices.

The success of the GAN-based framework, scDEC-Hi-C, proposed by Liu et al. [67], in analyzing single-cell Hi-C data can be primarily attributed to the model's capacity to handle sparse and highly variable data. The challenges posed by single-cell Hi-C data, including sparsity and cell-to-cell variability, are effectively mitigated by the synthetic data generated by the GAN. The generator learns to produce synthetic data that mimic the distribution of actual single-cell Hi-C data, effectively augmenting the sparse dataset. Simultaneously, the discriminator ensures the quality of the synthetic data, leading to enhanced understanding of chromatin architecture variability across cell types and superior clustering and imputation performance.

7.2. Trends, Challenges, and Future Directions in Recent Studies

The application of GANs for enhancing the resolution of Hi-C data represents a growing trend in genomics, as showcased in the works by Hu et al. [64], Hong et al. [65], and Liu et al. [66]. Despite the impressive advancements, several challenges and opportunities for future research stand out.

The successful development and application of EnHiC [64] and DeepHiC [65] have highlighted the effectiveness of GANs in enhancing the resolution of Hi-C matrices from low-resolution and low-coverage sequencing data. However, these methods rely on the availability of high-quality, high-resolution Hi-C data for training, which may not always be accessible. Future research should therefore aim to develop techniques that allow for efficient training of GANs with limited or noisy high-resolution data.

Moreover, although the hicGAN model [66] has shown promise in inferring high-resolution Hi-C data from low-resolution counterparts, the issue of verifying the authenticity of generated Hi-C matrices remains. Consequently, an important direction for future studies would be to establish rigorous evaluation metrics and validation methods to ensure the biological validity and relevance of GAN-generated Hi-C data.

Another noteworthy trend is the application of GANs to single-cell Hi-C data, exemplified by the scDEC-Hi-C model [67]. As single-cell technologies continue to advance, the application of GANs in this domain offers immense opportunities for uncovering cell-to-cell variability in chromatin architecture. However, dealing with the high level of noise and data sparsity inherent to single-cell Hi-C data presents a major challenge. Future research should focus on improving the robustness of GANs to these issues and developing methods for efficient handling of large-scale single-cell Hi-C datasets.

8. Other Studies Using Generative Adversarial Network for Gene Expression Data

The application of GANs in the domain of gene expression data extends beyond the topics explored in the preceding sections. Numerous studies have harnessed the potential of GANs to tackle other intricate problems in this domain, broadening the scope and utility of GANs in bioinformatics.

These varied studies can be categorized by the specific challenge in gene expression data that they aim to address using GANs. For instance, some investigations focus on the improvement of existing bioinformatics algorithms using synthetic gene expression data generated by GANs. In these cases, GANs are used as a tool to refine the inputs of other models, effectively enhancing their performance by enriching the training data or by adding complexity to the model training process. Other studies apply GANs as a mechanism for data imputation, seeking to fill in the gaps or resolve the inconsistencies within gene expression datasets. These applications underscore the ability of GANs to model intricate data distributions and generate realistic, high-dimensional synthetic data. Several investigations also exploit GANs to facilitate the interpretation of gene expression data. By generating synthetic data that closely mirrors real-world distributions, these studies use GANs to tease out the latent structures within the data, subsequently revealing novel biological insights.

Additionally, there are studies where GANs are employed as a feature selection tool. In these cases, GANs are used to reduce the high dimensionality of gene expression data, helping to identify the most informative genes or gene clusters for specific biological phenomena or conditions. Table 9 showcases a collection of recent studies.

Table 9. Summary of other recent studies using generative adversarial networks in gene expression data.

Author	Contributions
Booker et al. [68]	Used GANs to learn and replicate the distribution of population genetic alignments across evolutionary histories.
Salekin et al. [69]	Introduced MR-GAN for predicting epitranscriptome modifications sites from transcript sequences.
Pati et al. [70]	Presented Sim-GAN for estimating missing values in microarray data based on the similarity index.
Yuan et al. [71]	Proposed MSFF-CDCGAN, a GAN-based method for predicting RNA secondary structure.
Lee et al. [72]	Utilized GAN to explore the role of the OASIS family in defense mechanisms against cocaine-induced hypoxia and ischemia (HI).
Kim et al. [73]	Developed ICEGAN for network estimation applied to gene expression data.
Tsourtis et al. [74]	Presented a semi-interpretable generative model using Renyi divergence for gene expression data.
Uthamacumaran [75]	Applied GANs with Waddington landscape reconstruction and fractal dimension analysis to infer dynamics in glioblastoma-derived stem cells.
Xue et al. [76]	Used deep generative models to learn cellular states underlying transcriptomic responses to perturbations.
Qiao et al. [77]	Introduced iPro-GAN for identifying gene promoters and their strength.
Han et al. [78]	Proposed Gene-CWGAN, a data enhancement method for gene expression profiles.
Park et al. [79]	Applied GANs to predict the molecular progression of Alzheimer’s disease using RNA-seq data.
Vinas et al. [80]	Introduced PMI and GAIN-GTEx, deep learning methods for gene expression imputation.
Sui et al. [81]	Proposed a deep learning-based radiogenomic framework to link lung tumor images and genomic data.

8.1. Recent Studies: 2019–2023

The effective application of GANs in the domain of gene expression data can be attributed to several key principles and mechanisms inherent in their design. Firstly, GANs are adept at modeling complex, high-dimensional data distributions, a feature that is especially useful for handling the intricate nature of gene expression data. Furthermore, GANs employ a unique training procedure where two neural networks, i.e., the generator and the discriminator, compete in a game-theoretic framework. This adversarial setup, combined with the ability to learn latent representations from the data, allows GANs to generate synthetic data that is both realistic and closely mirrors the original distribution. These attributes largely contribute to the successful deployment of GANs in various bioinformatics applications, as showcased in the studies discussed below.

Booker et al. [68] successfully utilized a Deep-Convolutional Wasserstein GAN to model the distribution of population genetic alignments. The Wasserstein GAN variant was chosen due to its improved stability in training and capability to handle mode collapse, a common problem in GAN training. Moreover, the deep convolutional architecture was instrumental in capturing local dependencies and spatial correlations in the alignment data. These factors contributed to the model’s ability to generate synthetic examples resembling the original data, including key characteristics such as site-frequency spectrum, population differentiation, and linkage disequilibrium patterns.

Lee et al. [72] applied GANs to analyze the function of the OASIS family in combating cocaine-induced hypoxia and ischemia. The strength of GANs lies in their capacity to learn a target distribution in an unsupervised manner. In this study, the GAN’s ability to model the underlying distribution of the gene expression data enabled the discovery of novel patterns and associations linked to the biological phenomena under investigation. This, in turn, shed light on the role of the OASIS family in hypoxia and ischemia induced by cocaine.

Pati et al. [70] addressed missing values in microarray data using Sim-GAN. The essence of the method’s success lies in the unique approach of combining GANs with similarity indices. This combination capitalizes on the GAN’s ability to learn and generate

realistic data, and the similarity index's function of identifying similar patterns within the data. Consequently, the generated values closely resemble the true values, thus enhancing the utility and interpretability of the imputed gene expression datasets.

Salekin et al. [69] developed MR-GAN to predict epitranscriptome modification sites from transcript sequences. The success of this model is grounded in the inherent ability of GANs to capture intricate patterns within high-dimensional data. Specifically, the model leverages the GAN's capacity for deep representation learning to identify key sequence features that determine modification sites.

The study by Yuan et al. [71] demonstrates another unique application of GANs, in which RNA data are translated into grayscale images to predict RNA secondary structures using MSFF-CDCGAN. The translation into a different data representation medium allowed the GAN to exploit spatial correlations within the data, thereby improving the model's ability to learn and generate realistic secondary RNA structures.

8.2. Trends, Challenges, and Future Directions in Recent Studies

The utilization of GANs has touched upon various facets of bioinformatics, from improving data quality [70,80] to novel applications in disease progression study [75,79] and bioinformatic data augmentation and prediction [69,71,73,78]. The continued development and exploration of GANs herald an exciting future for gene expression data analysis.

However, alongside this optimism come significant challenges. One such challenge is the requirement for large amounts of high-quality data for GANs to perform optimally. For instance, the models developed by Pati et al. [70] and Vinas et al. [80] rely heavily on the completeness and accuracy of the microarray and RNA-seq data, respectively. If these data sets are not meticulously curated and preprocessed, the models' performance could be compromised, leading to less reliable outcomes.

Furthermore, while GANs are potent tools for data augmentation and prediction [69,71,73,78], interpreting the generated features and understanding the complex internal processes of the models remain considerable challenges. The models are often treated as black boxes due to their intricate structure, which makes it difficult to explain the underlying reasoning and decision-making processes. This opacity can inhibit the broader acceptance of GANs, particularly in clinical settings where interpretability is highly valued.

The vast potential and aforementioned challenges usher in several directions for future research. Firstly, the development of methods that require less data or that can effectively handle noisy or incomplete data is highly desirable. This would reduce the dependency on large, high-quality datasets and potentially extend the application of GANs to areas where such data are not readily available. Secondly, advances in the interpretability and transparency of GANs could enhance their acceptance and usability. This could involve the development of novel techniques to interpret the internal workings of the models, or potentially integrating GANs with more interpretable machine learning approaches.

As the studies of Lee et al. [72], Park et al. [79], and Uthamacumaran [75] suggest, there is a growing interest in applying GANs to understand complex biological processes and disease progression. Future work could involve the exploration of GANs in other diseases, or the integration of GANs with other data types, such as imaging or clinical data, to provide a more comprehensive understanding of disease processes. As the field continues to evolve, GANs are set to become an increasingly important tool in the analysis of gene expression data.

9. Discussion

Drawing upon the extensive research that has been conducted on GANs and their applications in gene expression data, it is clear that these tools have the potential to revolutionize the field of genomics. The remarkable adaptability of GANs, made possible through advancements in deep learning, has enabled their use in numerous areas of genomics, from improving the quality of gene expression data to facilitating more nuanced understanding of complex genomic structures.

The emergence of deep learning, a subset of machine learning, has been particularly transformative in this regard. In the domain of deep learning, a network learns from the data by adjusting its internal parameters [91–94]. In this way, the model is able to perform complex tasks such as complex games with reinforcement learning [95–97] and, as we have explored in this review, gene expression data analysis. An essential component of deep learning models is the activation function, which introduces non-linearity into the model and determines whether a particular neuron should be activated.

While the results to date have been promising, challenges persist. A significant barrier lies in the black box nature of GANs, as their complex inner workings often lack transparency. This can lead to skepticism, particularly in clinical settings where understanding the decision-making process is crucial. Another hurdle is the dependence of GANs on large, high-quality datasets for optimal performance. The availability of such data is not always guaranteed, which could potentially impact the efficacy of these models. Moreover, the existing methods for training GANs with limited or noisy data are still in the early stages, necessitating further research and development.

Moving forward, an emphasis on the development of more robust and versatile GAN models will be paramount. Efforts should be made to improve their ability to handle different types of genetic data, and specific focus should be directed towards enhancing their performance with limited or low-quality data. Furthermore, in-depth exploration into the interpretability of these models will be essential to garner wider acceptance of their usage. The rapid progress in this field is encouraging, and the potential applications of GANs in genomics are expansive. With continued exploration and optimization, it is anticipated that GANs will play an increasingly pivotal role in the advancement of genomics, bringing us closer to realizing the full potential of gene expression data analysis.

10. Conclusions

The remarkable ascent of GANs within the scope of gene expression data has undeniably transformed the landscape of computational genomics. It has been demonstrably elucidated throughout this review that GANs play an instrumental role in solving complex problems across an array of genomics applications, from synthetic gene expression data generation and cancer diagnosis to single-cell RNA-seq analysis and genome analysis.

We are at an inflection point in the genomics era, where we have both the computational tools and the demand to generate synthetic gene expression data. GANs, with their unique abilities, have shown immense potential in meeting this demand, unlocking new avenues for exploratory research in genomics and gene expression studies. They have demonstrated a considerable capability to enrich data sets, mitigate data imbalance issues, and bolster the robustness of downstream analytical tasks, all while upholding ethical and logistical considerations in human subject research.

This review's paramount importance lies in its contribution to the swiftly evolving domain of GANs in genomics. Given the brisk pace at which deep learning technologies are advancing, this review plays a crucial role in digesting, summarizing, and synthesizing the latest developments, thereby equipping researchers with current and pertinent knowledge. This, in turn, fuels further exploration and innovation in the field, driving the wheel of progress forward.

The rapid advancements in GANs necessitate continued investigation and critical appraisal of their impact on gene expression data. By providing a comprehensive, accessible, and up-to-date overview of the field, this review serves as a stepping stone for future research. We fervently hope this exploration will serve as a beacon, guiding further research and fostering the evolution of generative artificial intelligence within the domain of genomics.

Funding: This work was supported by a research grant funded by Generative Artificial Intelligence System Inc. (GAIS).

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: In the preparation of this paper, the author utilized Grammarly, an AI-assisted language editing tool, to enhance the grammatical correctness and fluency of our sentences.

Conflicts of Interest: Minhyeok Lee has received research grants from Generative Artificial Intelligence System Inc. The funding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. OpenAI Technical Report. 2019. Available online: https://d4mucfpksyvv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 15 May 2023).
- 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.
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. OpenAI Technical Report. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 15 May 2023).
- Lee, M. A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics* **2023**, *11*, 2320. [CrossRef]
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
- Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
- Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
- Yeom, T.; Lee, M. DuDGAN: Improving Class-Conditional GANs via Dual-Diffusion. *arXiv* **2023**, arXiv:2305.14849.
- Jabbar, A.; Li, X.; Omar, B. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–49. [CrossRef]
- Aggarwal, A.; Mittal, M.; Battineni, G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100004. [CrossRef]
- Ko, K.; Lee, M. ZIGNeRF: Zero-shot 3D Scene Representation with Invertible Generative Neural Radiance Fields. *arXiv* **2023**, arXiv:2306.02741.
- Yinka-Banjo, C.; Ugot, O.A. A review of generative adversarial networks and its application in cybersecurity. *Artif. Intell. Rev.* **2020**, *53*, 1721–1736. [CrossRef]
- Cai, Z.; Xiong, Z.; Xu, H.; Wang, P.; Li, W.; Pan, Y. Generative adversarial networks: A survey toward private and secure applications. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–38. [CrossRef]
- Chen, Y.; Yang, X.H.; Wei, Z.; Heidari, A.A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q. Generative adversarial networks in medical image augmentation: A review. *Comput. Biol. Med.* **2022**, *54*, 105382. [CrossRef]
- Lu, Y.; Chen, D.; Olaniyi, E.; Huang, Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput. Electron. Agric.* **2022**, *200*, 107208. [CrossRef]
- Singh, N.K.; Raza, K. Medical image generation using generative adversarial networks: A review. In *Health Informatics: A Computational Perspective in Healthcare*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 77–96.
- Ko, K.; Yeom, T.; Lee, M. Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains. *Neural Netw.* **2023**, *162*, 330–339. [CrossRef] [PubMed]
- 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]
- Wang, L.; Chen, W.; Yang, W.; Bi, F.; Yu, F.R. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access* **2020**, *8*, 63514–63537. [CrossRef]
- Lan, L.; You, L.; Zhang, Z.; Fan, Z.; Zhao, W.; Zeng, N.; Chen, Y.; Zhou, X. Generative adversarial networks and its applications in biomedical informatics. *Front. Public Health* **2020**, *8*, 164. [CrossRef]
- Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3313–3332. [CrossRef]
- Buccitelli, C.; Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **2020**, *21*, 630–644. [CrossRef]

24. Yu, H.; Welch, J.D. MichiGAN: Sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome Biol.* **2021**, *22*, 158. [\[CrossRef\]](#)
25. Yelmen, B.; Decelle, A.; Ongaro, L.; Marnetto, D.; Tallec, C.; Montinaro, F.; Furtlehner, C.; Pagani, L.; Jay, F. Creating artificial human genomes using generative neural networks. *PLoS Genet.* **2021**, *17*, e1009303. [\[CrossRef\]](#)
26. Hazra, D.; Kim, M.R.; Byun, Y.C. Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome. *Int. J. Mol. Sci.* **2022**, *23*, 3701. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Zrimec, J.; Fu, X.; Muhammad, A.S.; Skrekas, C.; Jauniskis, V.; Speicher, N.K.; Boerlin, C.S.; Verendel, V.; Chehreghani, M.H.; Dubhashi, D.; et al. Controlling gene expression with deep generative design of regulatory DNA. *Nat. Commun.* **2022**, *13*, 5099. [\[CrossRef\]](#)
28. Ahmed, K.T.; Sun, J.; Cheng, S.; Yong, J.; Zhang, W. Multi-omics data integration by generative adversarial network. *Bioinformatics* **2022**, *38*, 179–186. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Vinas, R.; Andres-Terre, H.; Lio, P.; Bryson, K. Adversarial generation of gene expression data. *Bioinformatics* **2022**, *38*, 730–737. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Marouf, M.; Machart, P.; Bansal, V.; Kilian, C.; Magruder, D.S.; Krebs, C.F.; Bonn, S. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **2020**, *11*, 166. [\[CrossRef\]](#)
31. Chaudhari, P.; Agrawal, H.; Kotecha, K. Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Comput.* **2020**, *24*, 11381–11391. [\[CrossRef\]](#)
32. Kwon, C.; Park, S.; Ko, S.; Ahn, J. Increasing prediction accuracy of pathogenic staging by sample augmentation with a GAN. *PLoS ONE* **2021**, *16*, e0250458. [\[CrossRef\]](#)
33. Mendez-Lucio, O.; Baillif, B.; Clevert, D.A.; Rouquie, D.; Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **2020**, *11*, 10. [\[CrossRef\]](#)
34. Chen, X.; Roberts, R.; Tong, W.; Liu, Z. Tox-GAN: An Artificial Intelligence Approach Alternative to Animal Studies—A Case Study with Toxicogenomics. *Toxicol. Sci.* **2022**, *186*, 242–259. [\[CrossRef\]](#)
35. Tamilmani, G.; Devi, V.B.; Sujithra, T.; Shajin, F.H.; Rajesh, P. Cancer MiRNA biomarker classification based on Improved Generative Adversarial Network optimized with Mayfly Optimization Algorithm. *Biomed. Signal Process. Control* **2022**, *75*, 103545. [\[CrossRef\]](#)
36. Xiao, Y.; Wu, J.; Lin, Z. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Comput. Biol. Med.* **2021**, *135*, 104540. [\[CrossRef\]](#)
37. Wang, Y.; Zhang, S. Prediction of Tumor Lymph Node Metastasis Using Wasserstein Distance-Based Generative Adversarial Networks Combining with Neural Architecture Search for Predicting. *Mathematics* **2023**, *11*, 729. [\[CrossRef\]](#)
38. Bi, X.A.; Mao, Y.; Luo, S.; Wu, H.; Zhang, L.; Luo, X.; Xu, L. A novel generation adversarial network framework with characteristics aggregation and diffusion for brain disease classification and feature selection. *Briefings Bioinform.* **2022**, *23*, bbac454. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Targonski, C.; Bender, M.R.; Shealy, B.T.; Husain, B.; Paseman, B.; Smith, M.C.; Feltus, F.A. Cellular State Transformations Using Deep Learning for Precision Medicine Applications. *Patterns* **2020**, *1*, 6. [\[CrossRef\]](#)
40. Park, C.; Oh, I.; Choi, J.; Ko, S.; Ahn, J. Improved Prediction of Cancer Outcome Using Graph-Embedded Generative Adversarial Networks. *IEEE Access* **2021**, *9*, 20076–20088. [\[CrossRef\]](#)
41. Yang, Q.; Li, X. BiGAN: LncRNA-disease association prediction based on bidirectional generative adversarial network. *BMC Bioinform.* **2021**, *22*, 1–17. [\[CrossRef\]](#)
42. Gutta, C.; Morhard, C.; Rehm, M. Applying a GAN-based classifier to improve transcriptome-based prognostication in breast cancer. *PLoS Comput. Biol.* **2023**, *19*, e1011035. [\[CrossRef\]](#)
43. Mansoor, M.; Nauman, M.; Rehman, H.U.; Benso, A. Gene Ontology GAN (GOGAN): A novel architecture for protein function prediction. *Soft Comput.* **2022**, *26*, 7653–7667. [\[CrossRef\]](#)
44. Bahrami, M.; Maitra, M.; Nagy, C.; Turecki, G.; Rabiee, H.R.; Li, Y. Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics* **2021**, *37*, 1345–1351. [\[CrossRef\]](#)
45. Jeon, M.; Xie, Z.; Evangelista, J.E.; Wojciechowicz, M.L.; Clarke, D.J.B.; Ma’ayan, A. Transforming L1000 profiles to RNA-seq-like profiles with deep learning. *BMC Bioinform.* **2022**, *23*, 374. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Wang, D.; Hou, S.; Zhang, L.; Wang, X.; Liu, B.; Zhang, Z. iMAP: Integration of multiple single-cell datasets by adversarial paired transfer networks. *Genome Biol.* **2021**, *22*, 63. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Xu, Y.; Zhang, Z.; You, L.; Liu, J.; Fan, Z.; Zhou, X. scIGANs: Single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res.* **2020**, *48*, e85. [\[CrossRef\]](#)
48. Wang, X.; Zhang, C.; Zhang, Y.; Meng, X.; Zhang, Z.; Shi, X.; Song, T. IMGG: Integrating Multiple Single-Cell Datasets through Connected Graphs and Generative Adversarial Networks. *Int. J. Mol. Sci.* **2022**, *23*, 2082. [\[CrossRef\]](#)
49. Lin, E.; Mukherjee, S.; Kannan, S. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinform.* **2020**, *21*, 64. [\[CrossRef\]](#)
50. Xu, Z.; Luo, J.; Xiong, Z. scSemiGAN: A single-cell semi-supervised annotation and dimensionality reduction framework based on generative adversarial network. *Bioinformatics* **2022**, *38*, 5042–5048. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Zhu, M.; Lai, Y. Improvements Achieved by Multiple Imputation for Single-Cell RNA-Seq Data in Clustering Analysis and Differential Expression Analysis. *J. Comput. Biol.* **2022**, *29*, 634–649. [\[CrossRef\]](#)

52. Ding, J.; Regev, A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. *Nat. Commun.* **2021**, *12*, 2554. [[CrossRef](#)] [[PubMed](#)]
53. Wei, X.; Dong, J.; Wang, F. scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation. *Bioinformatics* **2022**, *38*, 3377–3384. [[CrossRef](#)]
54. Reiman, D.; Manakkat Vijay, G.K.; Xu, H.; Sonin, A.; Chen, D.; Salomonis, N.; Singh, H.; Khan, A.A. Pseudocell Tracer-A method for inferring dynamic trajectories using scRNAseq and its application to B cells undergoing immunoglobulin class switch recombination. *PLoS Comput. Biol.* **2021**, *17*, e1008094. [[CrossRef](#)]
55. Wang, Y.; Liu, T.; Zhao, H. ResPAN: A powerful batch correction model for scRNA-seq data through residual adversarial networks. *Bioinformatics* **2022**, *38*, 3942–3949. [[CrossRef](#)]
56. Wang, Z.; Wang, J.; Kourakos, M.; Hoang, N.; Lee, H.H.; Mathieson, I.; Mathieson, S. Automatic inference of demographic parameters using generative adversarial networks. *Mol. Ecol. Resour.* **2021**, *21*, 2689–2705. [[CrossRef](#)]
57. Yang, H.; Gu, F.; Zhang, L.; Hua, X.S. Using generative adversarial networks for genome variant calling from low depth ONT sequencing data. *Sci. Rep.* **2022**, *12*, 8725. [[CrossRef](#)] [[PubMed](#)]
58. Kim, H.; Kim, Y.; Lee, C.Y.; Kim, D.G.; Cheon, M. Investigation of early molecular alterations in tauopathy with generative adversarial networks. *Sci. Rep.* **2023**, *13*, 732. [[CrossRef](#)]
59. Jiang, X.; Zhao, J.; Qian, W.; Song, W.; Lin, G.N. A Generative Adversarial Network Model for Disease Gene Prediction with RNA-seq Data. *IEEE Access* **2020**, *8*, 37352–37360. [[CrossRef](#)]
60. Lee, H.; Yeom, S.; Kim, S. BP-GAN: Interpretable Human Branchpoint Prediction Using Attentive Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 97851–97862. [[CrossRef](#)]
61. Du, B.; Tang, L.; Liu, L.; Zhou, W. Predicting LncRNA-Disease Association Based on Generative Adversarial Network. *Curr. Gene Ther.* **2022**, *22*, 144–151. [[CrossRef](#)]
62. Yan, X.; Wang, L.; You, Z.H.; Li, L.P.; Zheng, K. GANCDNA: A novel method for predicting circRNA-disease associations based on deep generative adversarial network. *Int. J. Data Min. Bioinform.* **2020**, *23*, 265–283. [[CrossRef](#)]
63. Wang, L.; Yan, X.; You, Z.H.; Zhou, X.; Li, H.Y.; Huang, Y.A. SGANRDA: Semi-supervised generative adversarial networks for predicting circRNA-disease associations. *Briefings Bioinform.* **2021**, *22*, bbab028. [[CrossRef](#)]
64. Hu, Y.; Ma, W. EnHiC: Learning fine-resolution Hi-C contact maps using a generative adversarial framework. *Bioinformatics* **2021**, *37*, 1272–1279. [[CrossRef](#)]
65. Hong, H.; Jiang, S.; Li, H.; Du, G.; Sun, Y.; Tao, H.; Quan, C.; Zhao, C.; Li, R.; Li, W.; et al. DeepHiC: A Generative Adversarial Network for Enhancing Hi-C Data Resolution. *PLoS Comput. Biol.* **2020**, *16*, e1007287. [[CrossRef](#)]
66. Liu, Q.; Lv, H.; Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **2019**, *35*, I99–I107. [[CrossRef](#)]
67. Liu, Q.; Zeng, W.; Zhang, W.; Wang, S.; Chen, H.; Jiang, R.; Zhou, M.; Zhang, S. Deep generative modeling and clustering of single cell Hi-C data. *Briefings Bioinform.* **2022**, *24*, bbac494. [[CrossRef](#)]
68. Booker, W.W.; Ray, D.D.; Schrider, D.R. This population does not exist: Learning the distribution of evolutionary histories with generative adversarial networks. *Genetics* **2023**, *224*, iyad063. [[CrossRef](#)] [[PubMed](#)]
69. Salekin, S.; Mostavi, M.; Chiu, Y.C.; Chen, Y.; Zhang, J.; Huang, Y. Predicting Sites of Epitranscriptome Modifications Using Unsupervised Representation Learning Based on Generative Adversarial Networks. *Front. Phys.* **2020**, *8*, 196. [[CrossRef](#)] [[PubMed](#)]
70. Pati, S.K.; Gupta, M.K.; Shai, R.; Banerjee, A.; Ghosh, A. Missing value estimation of microarray data using Sim-GAN. *Knowl. Inf. Syst.* **2022**, *64*, 2661–2687. [[CrossRef](#)]
71. Yuan, S.; Gong, Y.; Wang, G.; Zhang, B.; Liu, Y.; Zhang, H. MSFF-CDCGAN: A novel method to predict RNA secondary structure based on Generative Adversarial Network. *Methods* **2022**, *204*, 368–375. [[CrossRef](#)]
72. Lee, K.; Kim, T.; Cheon, M.; Yu, W. Unveiling OASIS family as a key player in hypoxia-ischemia cases induced by cocaine using generative adversarial networks. *Sci. Rep.* **2022**, *12*, 6734. [[CrossRef](#)] [[PubMed](#)]
73. Kim, I.; Lee, M.; Seok, J. ICEGAN: Inverse covariance estimating generative adversarial network. *Mach. Learn.-Sci. Technol.* **2023**, *4*, 025008. [[CrossRef](#)]
74. Tsourtis, A.; Papoutsoglou, G.; Pantazis, Y. GAN-Based Training of Semi-Interpretable Generators for Biological Data Interpolation and Augmentation. *Appl. Sci.* **2022**, *12*, 5434. [[CrossRef](#)]
75. Uthamacumaran, A. Pattern Detection on Glioblastoma's Waddington Landscape via Generative Adversarial Networks. *Cybern. Syst.* **2022**, *53*, 223–237. [[CrossRef](#)]
76. Xue, Y.; Ding, M.Q.; Lu, X. Learning to encode cellular responses to systematic perturbations with deep generative models. *NPJ Syst. Biol. Appl.* **2020**, *6*, 35. [[CrossRef](#)] [[PubMed](#)]
77. Qiao, H.; Zhang, S.; Xue, T.; Wang, J.; Wang, B. iPro-GAN: A novel model based on generative adversarial learning for identifying promoters and their strength. *Comput. Methods Programs Biomed.* **2022**, *215*, 106625. [[CrossRef](#)] [[PubMed](#)]
78. Han, F.; Zhu, S.; Ling, Q.; Han, H.; Li, H.; Guo, X.; Cao, J. Gene-CWGAN: A data enhancement method for gene expression profile based on improved CWGAN-GP. *Neural Comput. Appl.* **2022**, *34*, 16325–16339. [[CrossRef](#)]
79. Park, J.; Kim, H.; Kim, J.; Cheon, M. A practical application of generative adversarial networks for RNA-seq analysis to predict the molecular progress of Alzheimer's disease. *PLoS Comput. Biol.* **2020**, *16*, e1008099. [[CrossRef](#)] [[PubMed](#)]

80. Vinas, R.; Azevedo, T.; Gamazon, E.R.; Lio, P. Deep Learning Enables Fast and Accurate Imputation of Gene Expression. *Front. Genet.* **2021**, *12*, 624128. [[CrossRef](#)]
81. Sui, D.; Guo, M.; Ma, X.; Baptiste, J.; Zhang, L. Imaging Biomarkers and Gene Expression Data Correlation Framework for Lung Cancer Radiogenomics Analysis Based on Deep Learning. *IEEE Access* **2021**, *9*, 125247–125257. [[CrossRef](#)]
82. Li, R.; Li, L.; Xu, Y.; Yang, J. Machine learning meets omics: Applications and perspectives. *Briefings Bioinform.* **2022**, *23*, bbab460. [[CrossRef](#)]
83. Zottel, A.; Videtič Paska, A.; Jovčevska, I. Nanotechnology meets oncology: Nanomaterials in brain cancer research, diagnosis and therapy. *Materials* **2019**, *12*, 1588. [[CrossRef](#)]
84. González-Silva, L.; Quevedo, L.; Varela, I. Tumor functional heterogeneity unraveled by scRNA-seq technologies. *Trends Cancer* **2020**, *6*, 13–19. [[CrossRef](#)]
85. Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **2021**, *18*, 723–732. [[CrossRef](#)] [[PubMed](#)]
86. Pasquini, G.; Arias, J.E.R.; Schäfer, P.; Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 961–969. [[CrossRef](#)] [[PubMed](#)]
87. Yoon, J.; Jordon, J.; Schaar, M. Gain: Missing data imputation using generative adversarial nets. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 5689–5698.
88. Zhou, C.; McCarthy, S.A.; Durbin, R. YaHS: Yet another Hi-C scaffolding tool. *Bioinformatics* **2023**, *39*, btac808. [[CrossRef](#)]
89. Kong, S.; Zhang, Y. Deciphering Hi-C: From 3D genome to function. *Cell Biol. Toxicol.* **2019**, *35*, 15–32. [[CrossRef](#)] [[PubMed](#)]
90. Pal, K.; Forcato, M.; Ferrari, F. Hi-C analysis: From data generation to integration. *Biophys. Rev.* **2019**, *11*, 67–78. [[CrossRef](#)]
91. Lee, M. The Geometry of Feature Space in Deep Learning Models: A Holistic Perspective and Comprehensive Review. *Mathematics* **2023**, *11*, 2375. [[CrossRef](#)]
92. Lakshmana, K.; Kaluri, R.; Gundluru, N.; Alzamil, Z.S.; Rajput, D.S.; Khan, A.A.; Haq, M.A.; Alhussen, A. A review on deep learning techniques for IoT data. *Electronics* **2022**, *11*, 1604. [[CrossRef](#)]
93. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [[CrossRef](#)]
94. Lee, M. A Mathematical Interpretation of Autoregressive Generative Pre-Trained Transformer and Self-Supervised Learning. *Mathematics* **2023**, *11*, 2451. [[CrossRef](#)]
95. Matsuo, Y.; LeCun, Y.; Sahani, M.; Precup, D.; Silver, D.; Sugiyama, M.; Uchibe, E.; Morimoto, J. Deep learning, reinforcement learning, and world models. *Neural Netw.* **2022**, *152*, 267–275. [[CrossRef](#)]
96. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362*, 1140–1144. [[CrossRef](#)] [[PubMed](#)]
97. Lample, G.; Chaplot, D.S. Playing FPS games with deep reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

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.