International Journal of
Molecular Sciences

MDPI

Review

A Survey of Autoencoder Algorithms to Pave the Diagnosis of Rare Diseases

David Pratella [1], Samira Ait-El-Mkadem Saadi [2], Sylvie Bannwarth [2], Véronique Paquis-Fluckinger [2,†] and Silvia Bottini [1,*,†]

[1] Center of Modeling, Simulation and Interactions, Université Côte d'Azur, 06200 Nice, France; david.pratella@univ-cotedazur.fr
[2] Centre Hospitalier Universitaire (CHU) de Nice, Institute for Research on Cancer and Aging, Nice (IRCAN), Université Côte d'Azur, Inserm U1081, CNRS UMR 7284, 06200 Nice, France; saadi.s@chu-nice.fr (S.A.-E.-M.S.); bannwarthsylvie@yahoo.fr (S.B.); veronique.paquis@univ-cotedazur.fr (V.P.-F.)
* Correspondence: silvia.bottini@univ-cotedazur.fr; Tel.: +33-630566999
† Contributed as co-last author.

Abstract: Rare diseases (RDs) concern a broad range of disorders and can result from various origins. For a long time, the scientific community was unaware of RDs. Impressive progress has already been made for certain RDs; however, due to the lack of sufficient knowledge, many patients are not diagnosed. Nowadays, the advances in high-throughput sequencing technologies such as whole genome sequencing, single-cell and others, have boosted the understanding of RDs. To extract biological meaning using the data generated by these methods, different analysis techniques have been proposed, including machine learning algorithms. These methods have recently proven to be valuable in the medical field. Among such approaches, unsupervised learning methods via neural networks including autoencoders (AEs) or variational autoencoders (VAEs) have shown promising performances with applications on various type of data and in different contexts, from cancer to healthy patient tissues. In this review, we discuss how AEs and VAEs have been used in biomedical settings. Specifically, we discuss their current applications and the improvements achieved in diagnostic and survival of patients. We focus on the applications in the field of RDs, and we discuss how the employment of AEs and VAEs would enhance RD understanding and diagnosis.

Keywords: rare diseases; autoencoders; artificial intelligence; personalized medicine

Citation: Pratella, D.; Ait-El-Mkadem Saadi, S.; Bannwarth, S.; Paquis-Fluckinger, V.; Bottini, S. A Survey of Autoencoder Algorithms to Pave the Diagnosis of Rare Diseases. *Int. J. Mol. Sci.* **2021**, *22*, 10891. https://doi.org/10.3390/ijms221910891

Academic Editor: Emil Alexov

Received: 14 September 2021
Accepted: 7 October 2021
Published: 8 October 2021

1. Introduction

Genome regulation encompasses all facets of gene expression, from biochemical modifications of DNA to the physical arrangement of chromosomes and the activity of transcription mechanisms. Recently, several techniques have been developed to interrogate these complex processes in multiple dimensions (DNA, RNA, proteins, lipids, metabolites . . .), known as "omics". While these approaches can reveal physio-pathological mechanisms in the sample, the joint use of several omics on the same sample is key in the understanding of the associated phenotype [1].

The limited number of samples that can be collected are usually noisy, incompletely annotated, sparse, and high-dimensional (many variables), making it very challenging to develop integrative computational approaches with regard to this type of data. Nowadays, several machine learning approaches have been proposed to analyze multi-omics datasets [2]. Specifically, unsupervised approaches learn representations by identifying patterns in the data and extracting meaningful knowledge, while overcoming data complexities. Among such approaches, unsupervised learning methods via neural networks such as autoencoders (AEs) or variational autoencoders [3,4] (VAEs) have shown promising performances [5], with applications on various types of data, such as single-cell data [6],

multi-omics data [7], and metagenomics data [8] and in different contexts, such as cancer [9], bacterial infection [10] or in healthy patient tissues [11]. An AE learns a compressed representation (embedding) of the input data, passing the information through layers smaller than the previous one. The latent space will end up with a bottleneck layer composed of the most informative features of the original input data, and then will be used to reconstruct data in the most similar way. Through this process of compression, the algorithm will capture a better representation of the data structure (i.e., intrinsic relationships between the data variables), and therefore will allow for more accurate downstream analyses [12]. In this review, we will discuss about their usage in the field of rare diseases (RDs) and beyond, with a focus on why their implementation in such a context would be suitable in the near future.

1.1. RDs and Their Diagnosis

RDs are any disease that affects a small percentage of the population. In Europe, they affect less than 1 in 2000 citizens. There are more than 7000 RDs worldwide. Although individually rare, collectively, RDs are estimated to affect 350 million people globally. Most rare diseases are genetic and are present throughout a person's entire life, even if symptoms do not immediately appear. RDs are characterized by a wide diversity of symptoms, which can vary from patient to patient and can also appear to be similar to those of common diseases. These factors imply that RDs can often be misdiagnosed. According to the Global Genes organization, 8 out of 10 RDs are caused by a faulty gene, and approximately 75% affect children, yet it takes an average of 4.8 years to arrive at an accurate diagnosis. This is part of the reason for 30% of children with RDs not living to see their fifth birthday. There are numerous challenges and issues that need to be addressed, ranging from technical to theoretical aspects, such as the small number of patients, often children, the heterogeneity of the disease, and the limited amount of national/international data resources [13–15].

The development of new technologies, such as genomic analysis by means of next generation sequencing (NGS) and other omics technologies, has boosted the molecular understanding and diagnosis of RDs [16–24].

Despite a significant leap in the diagnostics of rare genetic diseases in recent years, more than half of patients with a suspected RD remain without a definite diagnosis [25]. Patients with RD who are not diagnosed or diagnosed late may experience a delay in the start of a specific treatment, which, in turn, could have irreversible consequences for their health, may prevent informed reproductive choice and could cause great stress for patients and their families.

1.2. Omics and Multi-Omics Approaches for RD Diagnosis

The development of high-throughput technologies in the past decade allowed us to generate a large amount of different data type, each of them representing different levels of information ranging from DNA level to protein level, including data such as genome, proteome, transcriptome, epigenome, metabolome [26,27]. All of these multi-omics data attempt to capture the biological machinery occurring in the living being, providing a high level of information. However, each technology individually cannot depict the entire biological complexity of most human diseases. The combination of multiple data types can compensate for missing or unreliable information in any single data type, and multiple sources of different biological measurements could point to the same results and low down the number of false positive [28].

The current challenge is to integrate these data together in order to decipher new levels of information that could be key in RD diagnosis, by identifying the causal mechanisms of those diseases. AEs and VAEs are very promising technologies to integrate and analyze data from different sources (e.g., multi-omics, patient registries, . . .) that can be used to overcome further challenges, such as low diagnostic rates, a reduced number of patients, and geographical dispersion.

2. Artificial Intelligence Methods for Biology

For several years, the various techniques of machine learning and deep learning have been widely used in image and visual recognition problems. The models developed so far can be classified into three popular categories, which are supervised, unsupervised and semi-supervised learning. For supervised methods, algorithms are fed with "labeled" data during a training step to infer a function which then will classify data or predict outcomes [29]. The purpose of unsupervised learning is to identify structures and features from a training dataset without the use of labeled data [30]. Most known unsupervised methods concern clustering algorithms such as hierarchical clustering or k-means clustering and dimensionality reduction methods such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [31] and uniform manifold approximation and projection (UMAP) [32]. Among unsupervised methods, artificial neural networks (ANN), particularly AEs and VAEs, have emerged as very promising methods to work with various biological problems and to integrate diverse types of data. Finally, semi-supervised learning is a learning problem that involves a small number of labeled examples and a large number of unlabeled examples. Learning problems of this type are challenging, as neither supervised nor unsupervised learning algorithms are able to make effective use of the mixtures of labeled and unlabeled data. As such, specialized semi-supervised learning algorithms are required. Although very promising, semi-supervised learning methods have mainly been applied for medical image analysis [33], which is out of the scope of the present review.

Basis of the AE Algorithm and Its Variant

AEs are composed of two main parts, which consist of an encoder and a decoder (Figure 1A). The encoder maps the highly dimensional input data into a latent variable consisting of one or multiple hidden layers of lower dimension. This bottleneck layer forces a compressed representation of the input data. The second part consists of the decoder, which attempts to reconstruct the input data from the embedding. The dimensionality reduction followed by the reconstruction of the input forces the model to only retain features with high variability, setting aside features with less variability. Autoencoders are often associated with the denoising procedure, because unimportant variations are automatically left out [34]. This loss is modeled through a loss function that considers the distance between compressed data and reconstructed data. The most commonly employed loss functions are mean squared error and Kullback–Leibler divergence.

Several variants of AEs have been proposed since they were first introduced. These variants mainly aim to address shortcomings, such as improved generalization, disentanglement, and modification to sequence input models. Some significant examples include the denoising autoencoder (DAE) [35] (Figure 1B), the sparse autoencoder (SAE) [36,37] (Figure 1C), and more recently the VAE [3,4] (Figure 1D). Each of these different generative models have their own specificity. The DAE takes as its input corrupted data for which some values have been randomly turned to zero. Usually, 50% of input nodes are set to zero; however, a lower percentage, around 30%, has been proposed [38]. This kind of AE has mainly been applied to images [35]. The SAE allows one to obtain a bottleneck layer without reducing the number of nodes in the hidden layers [36,37]. The loss function is defined using a sparsity penalty. This sparsity penalty can be defined by using the Kullback–Leibler divergence, which is a standard measure of the difference between two functions [36,37]. VAEs are probabilistic generative models, considering specific assumptions about the distribution of hidden layer features. They learn the true distribution of input features from latent variable distribution using the Bayesian approach and use stochastic inference to approximate a latent space defined by a mean "μ" and a standard deviation "σ" (Figure 1D). This property enables the possibility to compute probability distributions and thus generate new data [3,4]. The main difference between classical AE and VAE resides in the latent space which is continuous for the latter. These algorithms are scalable to large datasets and can deal with intractable posterior distributions by fit-

ting an approximate inference or recognition model, using a reparametrized variational lower bound estimator. They have been broadly tested and used for data compression or dimensionality reduction [11,39–44]. Their adaptability and potential to handle non-linear behavior have made them particularly well suited to work with complex data [7,9,45–48]. A recent benchmark proposed VAEs as the best-performing methods to detect cancer subtypes compared with other type of AE [7].

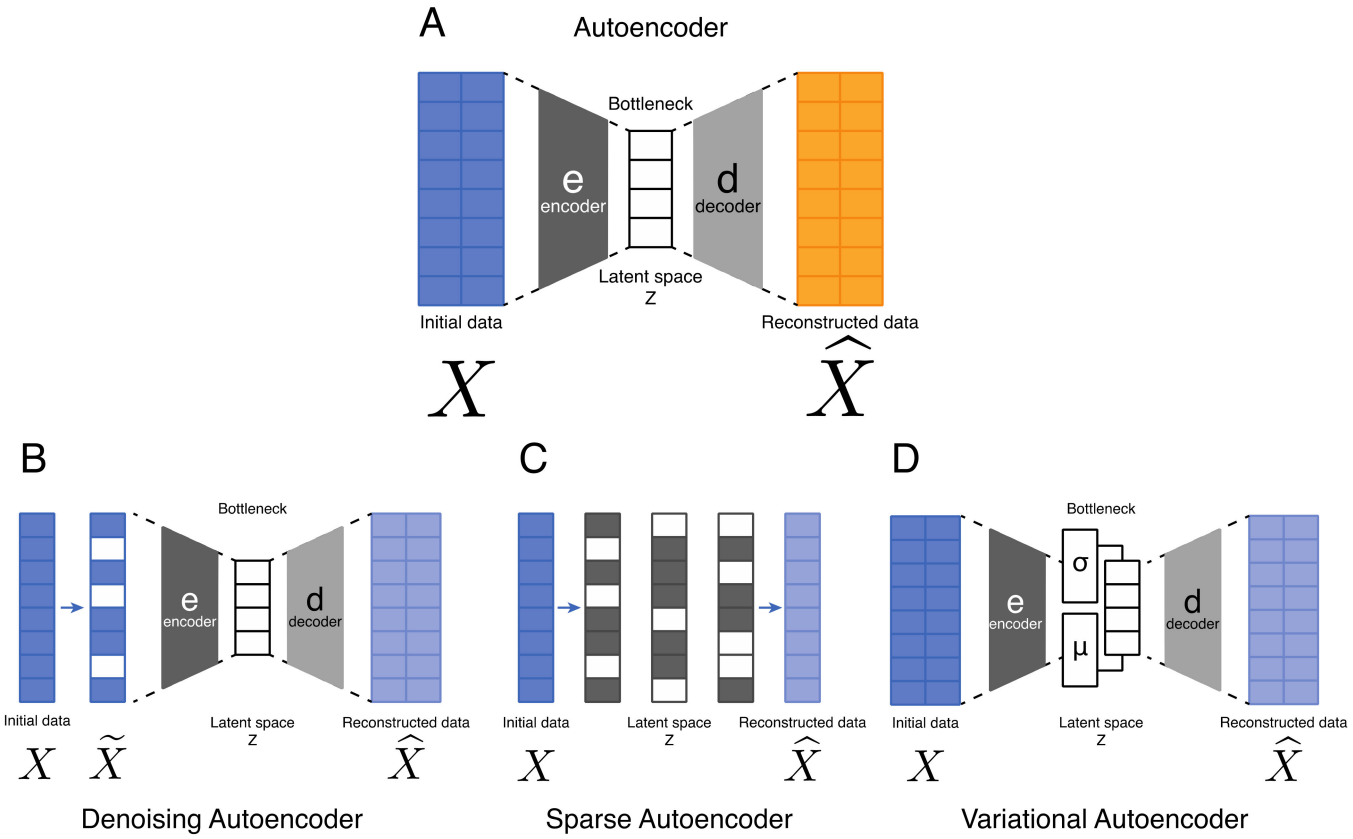

Figure 1. Different types of autoencoder. (**A**)—The classical autoencoder (AE) is composed of two main parts. First the encoder, annotated as "e", encodes the input data through a latent space "Z" by reducing the data dimensionality. The latent space corresponds to a vectorial space with a bottleneck constraint in order to force the algorithm to keep only the most variable features. The second part corresponds to the decoder, annotated as "d", that reconstructs the input data using the features encoded in the latent space. (**B**)—Denoising autoencoders (DAEs) are a category of AE where the input data are corrupted by setting nodes to a value of 0 (indicated in white). (**C**)—The sparse autoencoder (SAE) uses a penalty function. By penalizing the use of certain nodes (grey), these nodes are inactivated (white). Thus, the network is forced to learn features without reducing the number of nodes. (**D**)–The peculiarity of a variational autoencoder (VAE) is that the algorithm learns a distribution from the latent space "Z". This distribution is defined by a mean "μ" and a standard deviation "σ".

3. AE Applications in Biological and Medical Contexts beyond RD

The first applications of AE on biological data date from 2016. Tan et al. [49] developed ADAGE, a DAE (Figure 1B) to study microbe–host interactions. They showed that ADAGE is able to identify biological patterns and to extract meaningful features. By comparing their method with PCA and independent component analysis, they demonstrated that ADAGE was better at regrouping replicate samples and the biological features extracted by ADAGE were not clearly captured by other methods. They then improved it by constructing ensemble ADAGE [50] (eADAGE) by combining many individual ADAGE models into a single model. For each eADAGE model, they combined 100 models with identical parameters but distinct random seeds. Wang et al. [51] re-used the ADAGE package [49] to create a DAE model and used it on transcriptomic data from patients with lung cancer. They

identified a signature composed of 35 genes and concluded by proposing this signature as a novel diagnostic and prognostic biomarker for human lung adenocarcinoma. Chen et al. [52] used a SAE (Figure 1C) to study the transcriptomic machinery of yeast. This algorithm identified transcription factors with a fundamental role in yeast machinery by studying microarray gene expression. Furthermore, they found that SAE hidden layers correspond to common biological processes.

One example of the very first application of AE in the medical field is DeepPatient [53]. Taking advantage of electronic health records, Miotto et al. [53] developed a DAE-based method to improve clinical prediction for severe diabetes, schizophrenia and several cancers.

Finally, VAEs (Figure 1D) have been used in different biological contexts with different purposes, applied on different data type including proteomics, bulk RNA-seq and/or single-cell RNA-seq (scRNA-seq) and more. The origin of the data can vary, coming from healthy or diseased patients. Applications of AEs or VAEs on these data have been shown to improve downstream analysis and results, mainly for the identification of cell subtype, drug response prediction or multi-omics data integration.

Hereafter, we discuss some of the major applications of AE and its variants, in the biomedical field with an eye toward the advances in data analysis and developed algorithms (Supplementary Table S1, Figure 2).

3.1. AE Applications in Single Cell

Single-cell RNA sequencing (scRNA-seq) enables measurements of gene expression at the cell level and thus each of these cells will have its own transcriptome [54]. scRNA-seq allows to get a whole new level of information with more precision by comparison with bulk RNA-seq, where the sequencing results from a mixed cell population [55]. One of the main issues related to scRNA-seq data is the experimental noise that accompanies their generation. Indeed, at the single cell level, there is more variability in gene expression compared to an average cell population. Moreover, the low number of RNA transcripts available in single cell experiments will increase the rate of technical dropout events. This will provoke the scRNA-seq data to be highly sparse by including excessive zero counts that will cause the data to be zero-inflated, ending up with capturing only a small fraction of each cell transcriptome [56,57]. Recent research demonstrated the importance of correcting technical variation and showed improvement in downstream analysis [56–58]. One way to deal with this problem is to clean the data using a denoising algorithm [59]. Different works suggested several methods by using either a classical AE [34] or a VAE [6,11,39,40,60–62]. One solution is to go through an a priori modeling process. With VASC [39], the authors proposed to explicitly model the dropout events, which will help to find the non-linear hierarchical features representation of the original data. Using this approach, the authors asserted that VASC provided better dimension reduction and variational inference (scVI), both used a zero inflated negative binomial (ZINB) to model scRNA-seq noise. The ZINB allows one to take into account the RNA-seq count distribution, the overdispersion and the sparsity of the data by modeling the noise distribution in highly sparse count data. This causes the tool to learn gene-specific parameters such as the mean, dispersion and dropout probability and showed improvement in differential expression analysis, increasement in protein and RNA co-expression, enabling the discovery of subtle cellular phenotypes and increasing the correlation structure of key regulators [63].

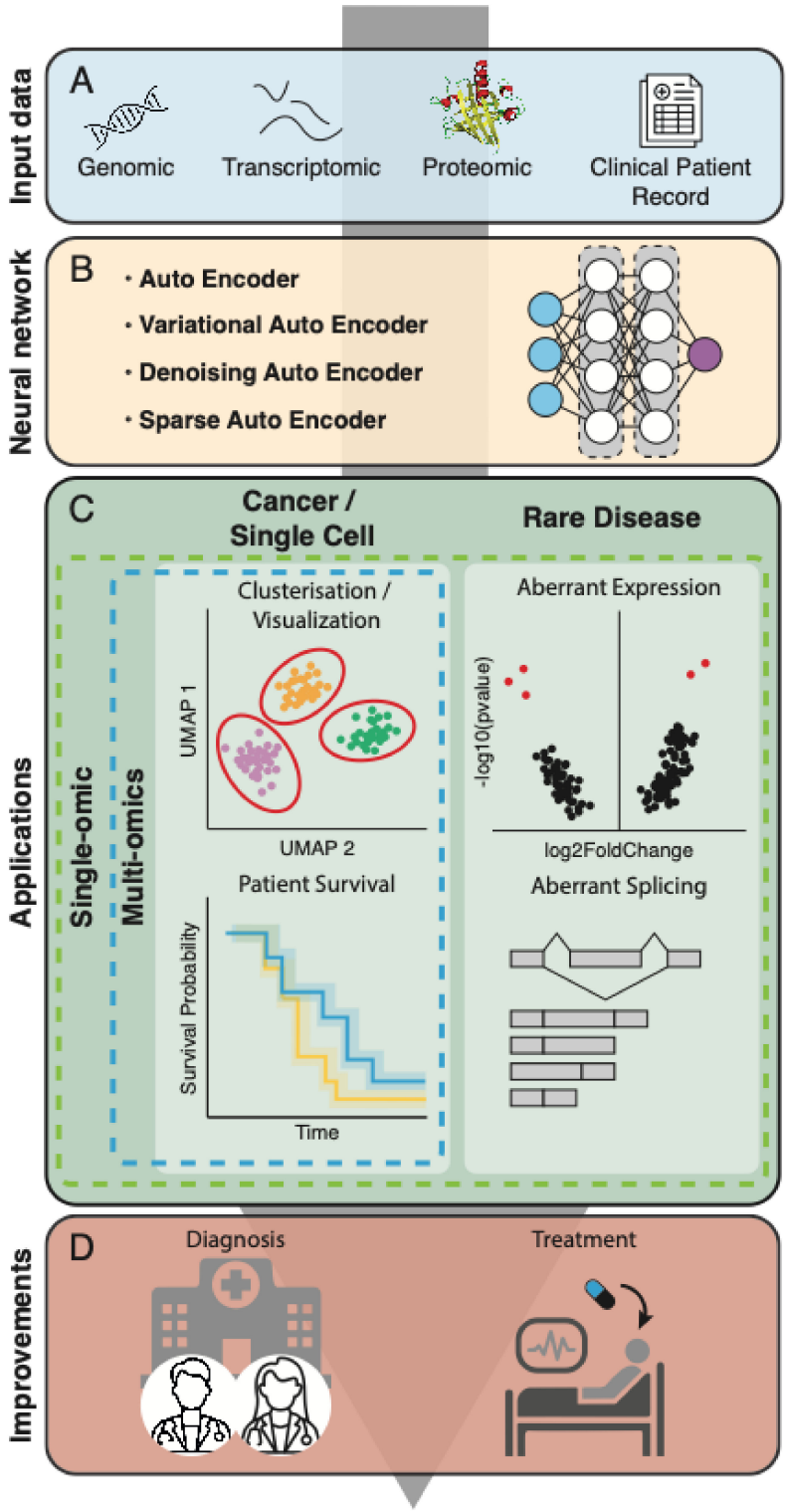

Figure 2. Autoencoders for personalized medicine approaches. (**A**)—Input omics data that can be of different types, such as omics, genomics, transcriptomics, proteomics, or clinal patient records. (**B**)—An AE-like algorithm is fed by input data as a single data type (single-omic) or multi data (multi-omics). (**C**)—Most common applications of these algorithms in the biomedical field and their achievements in terms of data analysis. Green dotted line groups AE applications on single-omic data, whereas blue dotted lines on multi-omics data. (**D**)—The results of the previous steps will enable improving patient diagnosis and treatment by providing powerful bioinformatics tools to physicians.

In addition to their denoising purpose, used for correcting the batch effect or dropout events, AEs can also be applied to single cell data for other tasks. For example, scGen [6] enables the prediction of events caused by an external perturbation due, for instance, to drugs or infection. Based on the association of a VAE and vector arithmetic, it models perturbation and infection response of cells across different cell types, studies and species. It works by learning cell-type and species-specific responses from features that distinguish responding from non-responding genes and cells. To demonstrate the performance of their tool, they applied it to the human Peripheral Blood Mononuclear Cells (PBMC) dataset [64] stimulated with interferon (IFN-β), showing good prediction on gene expression for stimulated CD4-T. They also evaluated scGen on data from Haber et al.'s [65] study, consisting of two datasets of intestinal epithelial cells impacted by *Salmonella* or *Heligmosomoides polygyrus* infection.

Another example of the use of VAE in single cell data is scVAE [11] and SCA [66], which are employed for classification/clustering tasks. scVAE uses different types of VAE with either a Gaussian or a Gaussian-mixture latent variable prior. It is able to obtain a higher Rand index, an index measuring the similarity between different clusters, showing better performances than Seurat [67], the state-of-the-art analysis tool for single cell data. On the other hand, SCA uses a SAE and showed good results for clustering single cell data, highlighting functional features. For example, the authors were able to identify genes highly involved in monocytes functionalities.

Semi-Supervised Generative Autoencoder (SISUA) [68] is a semi-supervised model based on the association of a VAE and CITE-seq (Cellular Indexing of Transcriptome and Epitopes by Sequencing) data. CITE-seq is a technique allowing researchers to obtain information from surface proteins. Because of the low amount of dropout in CITE-seq [69] data, the authors took advantage of this property to improve SC clustering results and notably obtained better separation between CD8 and CD4 proteins.

Recent technological advances have enabled simultaneous acquisitions of multiple omics data at the resolution of a single-cell, thus producing "multimodal" single-cell data. The first developed methods based on VAE for multi-omics analysis at single cell level were scMVAE [70] and totalVI [71]. These models have some limitations, including extensive preprocessing of data for training and latent variable interpretation difficulties. To overcome these limitations, Minoura et al. proposed scMM, a novel statistical framework for single-cell multi-omics analysis specialized in interpretable joint representation inference and predictions across modalities [72].

3.2. AE Applications in Cancer

Another application of AE on biological data concerns cancer data analyses. Several tools have been proposed with different strategies and different aims. Indeed, these methods focus either on drug response prediction [48,73,74] or on subtype cancer classification/stratification [7,47,51,73,75].

3.2.1. Drug Response Prediction

DeepDR [74] combines a network approach with AE. It is composed of three networks: i) a mutation encoder, ii) an expression encoder and iii) a drug response predictor network. The researchers showed that their tool performed better in drug response prediction compared to linear regression and SVM. The application of DeepDR revealed novel resistance mechanisms and drug targets. With the same aim, DeepProfile [73] and Dr.VAE [48] employ a VAE configuration. While DeepProfile uses a pre-trained VAE combined with a separately trained linear model to predict drug response, Dr.VAE is a semi-supervised method that learns a latent embedding of the gene expression used to feed a logistic regression classifier. The training data result from the combination of all the microarray datasets of the GEO database [76] for acute myeloid leukemia. Most of these methods outperformed currently used methods such as linear regression, SVM, PCA, k-means clustering [77].

3.2.2. Cancer Classification and Stratification

Cancer classification and stratification are fundamental to adapting the treatment depending on the cancer subtype and/or the prognostic, since cancer stage is closely related to cancer survival [78]. To address these tasks, different tools have been proposed. Tybalt et al. [75] proposed a VAE-based method learning features recapitulating tissues specific patterns. By training it on the cancer genome atlas (TCGA) dataset [79], the authors were able to identify different features such as patient sex, allowing classification, to compare melanoma tumors to other cancer type and to identify high-grade serous ovarian cancer (HGSC) subtypes. The stacked sparse auto-encoder (SSAE) is a semi-supervised deep learning strategy for cancer prediction using RNA-seq data [80]. This approach outperformed other methods for all three cancer data sets tested in various metrics.

Zhang et al. [47] proposed multi-omics data integration with an AE associated with K-means clustering to stratify high-risk neuroblastoma. They showed that their AE algorithm outperforms other non-AE methods such as iCluster [81] and PCA. Another method for multi-omics integration for cancer classification is OmiEmbed [82]. It combines the basic structure of VAE with a classifier to perform task-oriented feature extraction and multi-class classification. It yielded better performances than methods using only one type of omics data. With the same aim, Hira et al. [83] proposed the Maximum Mean Discrepancy VAE (MMD-VAE), which outperformed multi-omics analysis of ovarian cancer data.

To improve genomic functional characterization, Chen et al. [84] developed a gene superset autoencoder (GSAE), a multi-layer autoencoder model with the incorporation of a priori defined gene sets. They introduced the concept of the gene superset, an unbiased combination of gene sets, with weights trained by the AE, where each node in the latent layer is termed a superset, with the goal of determining the functional or clinical relevance of the learned gene supersets from the model.

Franco et al. [7] benchmarked four types of AE algorithms, including classic AE, DAE, SAE and VAE, to identify subtypes of cancer among glioblastoma multiforme, colon adenocarcinoma, kidney renal clear cell carcinoma and breast invasive carcinoma. They showed that even though AE performances varied depending on the dataset used, classical AE and VAE showed the best results, performing better than standard techniques for dimensionality reduction such as PCA, kernel PCA, and sparse PCA.

3.3. VAEs Structure for Data Integration

One application of VAE concerns multi-data integration [8,9,85], which is currently a challenge of high interest in computational biology. Several methods and configurations already exist [86], but without any clear consensus of the best one to use. Simidjievski et al. [9] proposed four different architectures for data integration using VAE structures: Variational Autoencoder with Concatenated Inputs (CNC-VAE), X-shaped Variational Autoencoder (X-VAE), Mixed-Modal Variational Autoencoder (MM-VAE) and Hierarchical Variational Autoencoder (H-VAE). By comparing the performances of these different methods, they showed that the H-VAE and X-VAE outperformed the other configurations, with a more stable behavior for the first one. The authors also suggested that data integration performances rely on data types, with some types being more amenable.

The study of Nissen et al. [8] proposed VAMB, a VAE to integrate two distinct data types. The first type is the sequence abundance defined by the individual number of reads mapped to each sequence. The second is the k-mer distribution, which corresponds to the number of substrings of length k contained in a sequence. By using these two types of data, they outperformed existing state-of-the-art tools. Another tool for data integration is deepDR [85], a network-based approach for studying drug repositioning where the authors integrated 10 different networks, including drug–disease, drug–side-effect, drug–target and drug–drug networks. By converting topological structure of each network into vector representation by using a random walk with restart algorithm, the authors were able to construct a positive point-wise mutual information (PPMI) matrix then fed to multimodal deep autoencoder (MDA) to concatenate all the different network. They extracted the

low-dimensional features from the middle layer of the MDA and then used it in a collective VAE (cVAE) to predict potential associations between drugs and diseases. By comparing their methods with baseline methods including random forest, kernelized Bayesian matrix factorization, support vector machine, random walk restart, they obtained better results.

One remarkable example of multi-omics data integration with AE is Multiview Factorization AutoEncoder (MAE) [87]. It combines multi-view learning, matrix factorization and AE with biological knowledge to integrate multi-omics data such as gene expression, DNA methylation and miRNA expression. The important contribution of this work is the introduction of external domain knowledge such as biological interaction networks to improve model generalizability and reduce the risk of overfitting. Another example of multi-omics integration through AE is Multiple Similarity Network Embedding (MSNE) [88]. MSNE integrates the multi-omics information by embedding the neighbor relations of samples defined by the random walk on multiple similarity networks. MSNE achieved outstanding performances for cancer subtyping compared to five other non-AE-based multi-omics integrative methods.

3.4. AE Applications in RDs

The development of omics technologies significantly improved the diagnosis of RDs. However, their success rate for detecting the responsible gene is far from complete. To fill this gap, the employment of RNA sequencing has been proposed as a complementary assay [89–93]. However, classical statistical methods have limited applications in the context of RD [94]. Consequently, there is an urgent need to develop novel computational approaches to resolve diagnostic deadlock and improve our knowledge of RD [1]. Brechtmann et al. developed OUTRIDER [95], an algorithm that uses an AE to model read-count expectations according to the gene covariation resulting from technical, environmental, or common genetic variations. The tool takes advantage of the generative model algorithm which reconstructing the RNA-seq data by fitting a negative binomial distribution and then computing a p-value and a Z-score. They used the Genotype Tissues Expression (GTEx) database [96], in which they injected simulated outliers in order to assess the sensitivity and the specificity of their tool. Additionally, the authors used data from Kremer et al.'s [89] study, consisting of individuals affected by rare mitochondrial disorders, with the goal to retrieve the aberrant gene expression manually identified and experimentally validated in the original publication.

Aberrant splicing is a major cause of rare disease. It is estimated that splicing mutations are responsible for 15–60% of human disease mutations [97–99]. By proposing FRASER [100], Mertes et al. responded to the lack of statistical significance assessments for splicing events in the field of RD. Their tool is based on a DAE and takes advantages of a beta binomial distribution, which takes overdispersion into account, and therefore is more suited for splicing events. To evaluate their tool, the authors used the same strategy employed for OUTRIDER. They injected splicing outliers in the GTEx and Kremer et al. datasets, and then computed a two-sided p-value along with a Z-score. To correct for multiple testing genome-wide, they used the FDR. FRASER showed better results compared to other methods, allowing them to identify several alternative splicing events including intron retention.

Although only few applications of AE and VAE has been developed in the context of RD, they have proven to be effective and have improved the diagnosis of RD. Thus, we foresee a rise of the employment of these technique in the field of RD.

4. Discussion and Conclusions: Open Challenges and Future Directions

With the recent advances in omics data production, we are able to perform various analyses. Omics enable us to enlarge the scope of biological data employed, enriching analysis and results. This progress has reduced the number of patients in diagnostic impasse, but it is still not enough. Multi-omics approaches are very promising for improving diagnostic performances, but several problems remain to be solved. Data analysis methods for

multi-omics are generally developed for cancer research, where large numbers of samples are available, which is not the case for RD. Therefore, there is a need to develop multi-omics approaches applicable to small cohorts. In this review, we comprehensively collected the basic but essential concepts and methods of AE, together with its recent applications in diverse biomedical studies (Figure 2). We have showed that the use of machine learning methods such as AE or VAE algorithms can improve analysis and results. However, few methods have been applied yet to RD. The identification of pathogenic events through measurement of aberrant gene expression levels is a very promising approach. With their tool based on an AE algorithm, Brechtmann et al. [95] successfully identified pathogenic genes candidates; however, the use of negative binomial distribution to model RNA-seq data limits the employability of the tool. To date, no methods have been proposed for multi-omics analysis in the field of RD. One of the challenges is the limited number of samples for RD with respect to other pathologies. The limited number of patients is not the only difficulty in applying existing algorithms for multi-omics integration to RD. These diseases are rare and heterogeneous, and the causative gene(s) are usually unique or "private" for each patient (or family). They require a methodology that identifies unique signatures, making it difficult to apply most of the multi-omics methods available because they are more suitable for identifying common signatures. AE proved to be useful in multi-omics data integration and could open the way to better-performing methods, especially in RD; however, they have some weaknesses [101]. They are highly sensitive to parameter tuning. It has been pointed out how the performances of each method could vary upon those hyperparameters. In their review, Hu and Green [101] proposed relying on independent third parties to benchmark and assess the different methods and tools. Extensive benchmarks are needed to learn more about AE and VAE performances in RD.

Despite the fact that the implementation of AE and VAE algorithms in RD is still in its infancy, it has opened the door to a more faithful understanding of the complex aspects of RD physiology, pathology, and treatment. Although much remains to be learned and developed, we believe that this review has captured the essence of this field and will enhance the use of AE in RD and inspire future breakthroughs in both the understanding and diagnosis of RD.

Supplementary Materials: https://www.mdpi.com/article/10.3390/ijms221910891/s1.

Author Contributions: Conceptualization, S.B. (Silvia Bottini) and V.P.-F.; writing—original draft preparation, D.P. and S.B. (Silvia Bottini); writing—review and editing, S.B. (Silvia Bottini), S.A.-E.-M.S. and S.B. (Sylvie Bannwarth); supervision, S.B. (Silvia Bottini); project administration, S.B. (Silvia Bottini) and V.P.-F.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) under reference number ANR15-IDEX-01.

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

References

1. Labory, J.; Fierville, M.; Ait-El-Mkadem, S.; Bannwarth, S.; Paquis-Flucklinger, V.; Bottini, S. Multi-Omics Approaches to Improve Mitochondrial Disease Diagnosis: Challenges, Advances, and Perspectives. *Front. Mol. Biosci.* **2020**, *7*, 590842. [CrossRef] [PubMed]
2. Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **2021**, *49*, 107739. [CrossRef]
3. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114. [cs, stat].
4. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv* **2014**, arXiv:1401.4082. [cs, stat].
5. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
6. Lotfollahi, M.; Wolf, F.A.; Theis, F.J. scGen predicts single-cell perturbation responses. *Nat. Methods* **2019**, *16*, 715–721. [CrossRef] [PubMed]

7. Franco, E.; Rana, P.; Cruz, A.; Calderón, V.; Azevedo, V.; Ramos, R.; Ghosh, P. Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. *Cancers* **2021**, *13*, 2013. [CrossRef] [PubMed]
8. Nissen, J.N.; Johansen, J.; Allesøe, R.L.; Sønderby, C.K.; Armenteros, J.J.A.; Grønbech, C.H.; Jensen, L.J.; Nielsen, H.B.; Petersen, T.N.; Winther, O.; et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **2021**, *39*, 555–560. [CrossRef]
9. Simidjievski, N.; Bodnar, C.; Tariq, I.; Scherer, P.; Andres-Terre, H.; Shams, Z.; Jamnik, M.; Liò, P. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Front. Genet.* **2019**, *10*, 1205. [CrossRef] [PubMed]
10. Deng, Y.; Bao, F.; Dai, Q.; Wu, L.F.; Altschuler, S.J. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat. Methods* **2019**, *16*, 311–314. [CrossRef]
11. Grønbech, C.H.; Vording, M.F.; Timshel, P.N.; Sønderby, C.K.; Pers, T.H.; Winther, O. scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **2020**, *36*, 4415–4422. [CrossRef]
12. Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
13. Christianson, A.; Howson, C.P.; Modell, B. March of Dimes: Global Report on Birth Defects, the Hidden Toll of Dying and Disabled Children. In *March of Dimes: Global Report on Birth Defects, the Hidden Toll of Dying and Disabled Children*; March of Dimes Birth Defects Foundation: Arlington, VA, USA, 2005.
14. Baird, P.A.; Anderson, T.W.; Newcombe, H.B.; Lowry, R.B. Genetic disorders in children and young adults: A population study. *Am. J. Hum. Genet.* **1988**, *42*, 677–693.
15. DI Resta, C.; Galbiati, S.; Carrera, P.; Ferrari, M. Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities. *EJIFCC* **2018**, *29*, 4–14.
16. Ng, S.B.; Turner, E.; Robertson, P.D.; Flygare, S.D.; Bigham, A.W.; Lee, C.; Shaffer, T.; Wong, M.; Bhattacharjee, A.; Eichler, E.E.; et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nat. Cell Biol.* **2009**, *461*, 272–276. [CrossRef] [PubMed]
17. Bamshad, M.J.; Ng, S.B.; Bigham, A.W.; Tabor, H.K.; Emond, M.J.; Nickerson, D.A.; Shendure, J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **2011**, *12*, 745–755. [CrossRef]
18. Ku, C.-S.; Naidoo, N.; Pawitan, Y. Revisiting Mendelian disorders through exome sequencing. *Qual. Life Res.* **2011**, *129*, 351–370. [CrossRef] [PubMed]
19. Boycott, K.M.; Vanstone, M.R.; Bulman, D.E.; MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* **2013**, *14*, 681–691. [CrossRef]
20. Shashi, V.; McConkie-Rosell, A.; Rosell, B.; Schoch, K.; Vellore, K.; McDonald, M.; Jiang, Y.-H.; Xie, P.; Need, A.; Goldstein, D.B. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet. Med.* **2013**, *16*, 176–182. [CrossRef]
21. Liew, W.K.M.; Ben-Omran, T.; Darras, B.T.; Prabhu, S.P.; De Vivo, D.C.; Vatta, M.; Yang, Y.; Eng, C.M.; Chung, W.K. Clinical Application of Whole-Exome Sequencing. *JAMA Neurol.* **2013**, *70*, 788–791. [CrossRef]
22. Yang, Y.; Muzny, D.M.; Reid, J.G.; Bainbridge, M.N.; Willis, A.; Ward, P.A.; Braxton, A.; Beuten, J.; Xia, F.; Niu, Z.; et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N. Engl. J. Med.* **2013**, *369*, 1502–1511. [CrossRef] [PubMed]
23. Lee, H.; Deignan, J.L.; Dorrani, N.; Strom, S.P.; Kantarci, S.; Quintero-Rivera, F.; Das, K.; Toy, T.; Harry, B.; Yourshaw, M.; et al. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **2014**, *312*, 1880–1887. [CrossRef]
24. Yang, Y.; Muzny, D.M.; Xia, F.; Niu, Z.; Person, R.; Ding, Y.; Ward, P.; Braxton, A.; Wang, M.; Buhay, C.; et al. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA* **2014**, *312*, 1870–1879. [CrossRef]
25. Clark, M.M.; Stark, Z.; Farnaes, L.; Tan, T.Y.; White, S.M.; Dimmock, D.; Kingsmore, S.F. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genom. Med.* **2018**, *3*, 16. [CrossRef] [PubMed]
26. Hasin, Y.; Seldin, M.; Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **2017**, *18*, 1–15. [CrossRef]
27. Beale, D.J.; Karpe, A.V.; Ahmed, W. Beyond Metabolomics: A Review of Multi-Omics-Based Approaches. In *Microbial Metabolomics*; Springer: Cham, Switzerland, 2016; pp. 289–312.
28. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef]
29. Cunningham, P.; Cord, M.; Delany, S.J. Supervised Learning. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*; Cord, M., Cunningham, P., Eds.; Cognitive Technologies; Springer: Berlin/Heidelberg, Germany, 2008; pp. 21–49. ISBN 9783540751717.
30. Greene, D.; Cunningham, P.; Mayer, R. Unsupervised Learning and Clustering. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*; Cord, M., Cunningham, P., Eds.; Cognitive Technologies; Springer: Berlin/Heidelberg, Germany, 2008; pp. 51–90. ISBN 9783540751717.
31. Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
32. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426. [cs, stat].
33. Cheplygina, V.; de Bruijne, M.; Pluim, J.P.W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal.* **2019**, *54*, 280–296. [CrossRef]

34. Eraslan, G.; Simon, L.M.; Mircea, M.; Mueller, N.S.; Theis, F.J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef]
35. Vincent, P.; LaRochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning-ICML '08, Helsinki, Finland, 5–9 July 2008; Association for Computing Machinery: New York, NY, USA; pp. 1096–1103.
36. Coates, A.; Ng, A.; Lee, H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
37. Makhzani, A.; Frey, B. K-Sparse Autoencoders. *arXiv* **2014**, arXiv:1312.5663. [cs].
38. Ferles, C.; Papanikolaou, Y.; Naidoo, K.J. Denoising Autoencoder Self-Organizing Map (DASOM). *Neural Networks* **2018**, *105*, 112–131. [CrossRef] [PubMed]
39. Wang, D.; Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genom. Proteom. Bioinform.* **2018**, *16*, 320–331. [CrossRef]
40. Ding, J.; Condon, A.; Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **2018**, *9*, 1–13. [CrossRef]
41. Gupta, A.; Wang, H.; Ganapathiraju, M. Learning structure in gene expression data using deep architectures, with an application to gene clustering. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 12–19 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1328–1335.
42. Amodio, M.; van Dijk, D.; Srinivasan, K.; Chen, W.S.; Mohsen, H.; Moon, K.R.; Campbell, A.; Zhao, Y.; Wang, X.; Venkataswamy, M.; et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **2019**, *16*, 1139–1145. [CrossRef] [PubMed]
43. Zhou, L.; Cai, C.; Gao, Y.; Su, S.; Wu, J. Variational Autoencoder for Low Bit-Rate Image Compression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2617–2620.
44. Tan, J.; Ung, M.; Cheng, C.; Greene, C.S. Unsupervised Feature Construction and Knowledge Extraction from Genome-Wide Assays of Breast Cancer with Denoising Autoencoders. *Pac. Symp. Biocomput.* **2015**, *20*, 132–143. [CrossRef]
45. Poirion, O.B.; Chaudhary, K.; Garmire, L.X. Deep Learning data integration for better risk stratification models of bladder cancer. *AMIA Jt. Summits Transl. Sci. Proc.* **2018**, *2017*, 197–206.
46. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **2018**, *24*, 1248–1259. [CrossRef] [PubMed]
47. Zhang, L.; Lv, C.; Jin, Y.; Cheng, G.; Fu, Y.; Yuan, D.; Tao, Y.; Guo, Y.; Ni, X.; Shi, T. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front. Genet.* **2018**, *9*, 477. [CrossRef]
48. Rampášek, L.; Hidru, D.; Smirnov, A.; Haibe-Kains, B.; Goldenberg, A. Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **2019**, *35*, 3743–3751. [CrossRef]
49. Tan, J.; Hammond, J.H.; Hogan, D.A.; Greene, C.S. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems* **2016**, *1*, 00025-15. [CrossRef] [PubMed]
50. Tan, J.; Doing, G.; Lewis, K.A.; Price, C.E.; Chen, K.M.; Cady, K.C.; Perchuk, B.; Laub, M.T.; Hogan, D.A.; Greene, C.S. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell Syst.* **2017**, *5*, 63–71. [CrossRef]
51. Wang, J.; Xie, X.; Shi, J.; He, W.; Chen, Q.; Chen, L.; Gu, W.; Zhou, T. Denoising Autoencoder, A Deep Learning Algorithm, Aids the Identification of a Novel Molecular Signature of Lung Adenocarcinoma. *Genom. Proteom. Bioinform.* **2020**, *18*, 468–480. [CrossRef] [PubMed]
52. Chen, L.; Cai, C.; Chen, V.; Lu, X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinform.* **2016**, *17*, 97–107. [CrossRef]
53. Miotto, R.; Li, L.; Kidd, B.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **2016**, *6*, 26094. [CrossRef]
54. Navin, N.; Kendall, J.; Troge, J.; Andrews, P.; Rodgers, L.; McIndoo, J.; Cook, K.; Stepansky, A.; Levy, D.; Esposito, D.; et al. Tumour evolution inferred by single-cell sequencing. *Nat. Cell Biol.* **2011**, *472*, 90–94. [CrossRef] [PubMed]
55. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
56. Brennecke, P.; Anders, S.; Kim, J.K.; Kolodziejczyk, A.; Zhang, X.; Proserpio, V.; Baying, B.; Benes, V.; Teichmann, S.; Marioni, J.; et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **2013**, *10*, 1093–1095. [CrossRef]
57. Kim, J.K.; Kolodziejczyk, A.; Illicic, T.; Teichmann, S.; Marioni, J.C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **2015**, *6*, 8687. [CrossRef]
58. Buettner, F.; Natarajan, K.N.; Casale, F.P.; Proserpio, V.; Scialdone, A.; Theis, F.J.; Teichmann, S.; Marioni, J.; Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **2015**, *33*, 155–160. [CrossRef]
59. Gomes, T.; Teichmann, S.A.; Talavera-López, C. Immunology Driven by Large-Scale Single-Cell Sequencing. *Trends Immunol.* **2019**, *40*, 1011–1021. [CrossRef] [PubMed]

60. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef]
61. Talwar, D.; Mongia, A.; Sengupta, D.; Majumdar, A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.* **2018**, *8*, 1–11. [CrossRef] [PubMed]
62. Lin, C.; Jain, S.; Kim, H.; Bar-Joseph, Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* **2017**, *45*, e156. [CrossRef] [PubMed]
63. Eraslan, G.; Avsec, Ž.; Gagneur, J.; Theis, F.J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403. [CrossRef]
64. Kang, H.M.; Subramaniam, M.; Targ, S.; Nguyen, M.; Maliskova, L.; McCarthy, E.; Wan, E.; Wong, S.; Byrnes, L.; Lanata, C.M.; et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **2018**, *36*, 89–94. [CrossRef]
65. Haber, A.L.; Biton, M.; Rogel, N.; Herbst, R.H.; Shekhar, K.; Smillie, C.; Burgin, G.; Delorey, T.M.; Howitt, M.R.; Katz, Y.; et al. A single-cell survey of the small intestinal epithelium. *Nature* **2017**, *551*, 333–339. [CrossRef]
66. Alessandri, L.; Cordero, F.; Beccuti, M.; Licheri, N.; Arigoni, M.; Olivero, M.; Di Renzo, M.F.; Sapino, A.; Calogero, R. Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining. *npj Syst. Biol. Appl.* **2021**, *7*, 1–10. [CrossRef]
67. Satija, R.; Farrell, J.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [CrossRef]
68. Trong, T.N.; Mehtonen, J.; González, G.; Kramer, R.; Hautamäki, V.; Heinäniemi, M. Semisupervised Generative Autoencoder for Single-Cell Data. *J. Comput. Biol.* **2020**, *27*, 1190–1203. [CrossRef]
69. Stoeckius, M.; Hafemeister, C.; Stephenson, W.; Houck-Loomis, B.; Chattopadhyay, P.K.; Swerdlow, H.; Satija, R.; Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **2017**, *14*, 865–868. [CrossRef]
70. Zuo, C.; Chen, L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings Bioinform.* **2020**, *22*. [CrossRef] [PubMed]
71. Gayoso, A.; Steier, Z.; Lopez, R.; Regier, J.; Nazor, K.L.; Streets, A.; Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **2021**, *18*, 272–282. [CrossRef] [PubMed]
72. Minoura, K.; Abe, K.; Nam, H.; Nishikawa, H.; Shimamura, T. ScMM: Mixture-of-Experts Multimodal Deep Generative Model for Single-Cell Multiomics Data Analysis. *bioRxiv* **2021**. [CrossRef]
73. Dincer, A.B.; Celik, S.; Hiranuma, N.; Lee, S.-I. DeepProfile: Deep Learning of Cancer Molecular Profiles for Precision Medicine. *bioRxiv* **2018**. [CrossRef]
74. Chiu, Y.-C.; Chen, H.-I.H.; Zhang, T.; Zhang, S.; Gorthi, A.; Wang, L.-J.; Huang, Y.; Chen, Y. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med. Genom.* **2019**, *12*, 18. [CrossRef]
75. Way, G.P.; Greene, C.S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* **2018**, *23*, 80–91. [CrossRef]
76. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef]
77. Baptista, D.; Ferreira, P.; Rocha, M. Deep learning for drug response prediction in cancer. *Brief. Bioinform.* **2021**, *22*, 360–379. [CrossRef]
78. Rami-Porta, R.; Crowley, J.; Goldstraw, P. Review the Revised TNM Staging System for Lung Cancer. *Ann. Thorac. Cardiovasc. Surg.* **2009**, *15*, 4–9.
79. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. Review the Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Współczesna Onkologia* **2015**, *1A*, 68–77. [CrossRef] [PubMed]
80. Xiao, Y.; Wu, J.; Lin, Z.; Zhao, X. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Comput. Methods Programs Biomed.* **2018**, *166*, 99–105. [CrossRef] [PubMed]
81. Shen, R.; Olshen, A.B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25*, 2906–2912. [CrossRef] [PubMed]
82. Zhang, X.; Xing, Y.; Sun, K.; Guo, Y. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. *Cancers* **2021**, *13*, 3047. [CrossRef] [PubMed]
83. Hira, M.T.; Razzaque, M.A.; Angione, C.; Scrivens, J.; Sawan, S.; Sarker, M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci. Rep.* **2021**, *11*, 1–16. [CrossRef]
84. Chen, H.-I.H.; Chiu, Y.-C.; Zhang, T.; Zhang, S.; Huang, Y.; Chen, Y. GSAE: An autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Syst. Biol.* **2018**, *12*, 142. [CrossRef]
85. Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **2019**, *35*, 5191–5198. [CrossRef]
86. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **2020**, *14*, 1177932219899051. [CrossRef]
87. Ma, T.; Zhang, A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genom.* **2019**, *20*, 944. [CrossRef] [PubMed]
88. Yuan, W.; He, K.; Shi, C.; Guan, D.; Tian, Y.; Al-Dhelaan, A.; Al-Dhelaan, M. Multi-view network embedding with node similarity ensemble. *World Wide Web* **2020**, *23*, 2699–2714. [CrossRef]

89. Kremer, L.S.; Bader, D.M.; Mertes, C.; Kopajtich, R.; Pichler, G.; Iuso, A.; Haack, T.B.; Graf, E.; Schwarzmayr, T.; Terrile, C.; et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **2017**, *8*, 15824. [CrossRef]
90. Cummings, B.B.; Marshall, J.L.; Tukiainen, T.; Lek, M.; Donkervoort, S.; Foley, A.R.; Bolduc, V.; Waddell, L.B.; Sandaradura, S.A.; O'Grady, G.L.; et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **2017**, *9*, eaal5209. [CrossRef]
91. Frésard, L.; Smail, C.; Ferraro, N.M.; Teran, N.A.; Li, X.; Smith, K.S.; Bonner, D.; Kernohan, K.D.; Marwaha, S.; Zappala, Z.; et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **2019**, *25*, 911–919. [CrossRef]
92. Gonorazky, H.D.; Naumenko, S.; Ramani, A.K.; Nelakuditi, V.; Mashouri, P.; Wang, P.; Kao, D.; Ohri, K.; Viththiyapaskaran, S.; Tarnopolsky, M.A.; et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* **2019**, *104*, 466–483. [CrossRef] [PubMed]
93. Lee, H.; Network, U.D.; Huang, A.Y.; Wang, L.-K.; Bs, A.J.Y.; Bs, G.R.; Eskin, A.; Ms, R.H.S.; Ms, N.D.; Bs, S.N.-R.; et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet. Med.* **2020**, *22*, 490–499. [CrossRef] [PubMed]
94. Schlieben, L.D.; Prokisch, H.; Yépez, V.A. How Machine Learning and Statistical Models Advance Molecular Diagnostics of Rare Disorders Via Analysis of RNA Sequencing Data. *Front. Mol. Biosci.* **2021**, *8*, 473. [CrossRef]
95. Brechtmann, F.; Mertes, C.; Matusevičiūtė, A.; Yépez, V.A.; Avsec, Ž.; Herzog, M.; Bader, D.M.; Prokisch, H.; Gagneur, J. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* **2018**, *103*, 907–917. [CrossRef] [PubMed]
96. The GTEx Consortium; Ardlie, K.G.; DeLuca, D.S.; Segre, A.V.; Sullivan, T.J.; Young, T.R.; Gelfand, E.T.; Trowbridge, C.A.; Maller, J.B.; Tukiainen, T.; et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648–660. [CrossRef]
97. Wang, G.-S.; Cooper, T.A. Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **2007**, *8*, 749–761. [CrossRef] [PubMed]
98. Park, E.; Pan, Z.; Zhang, Z.; Lin, L.; Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **2018**, *102*, 11–26. [CrossRef] [PubMed]
99. Taylor, K.; Sobczak, K. Intrinsic Regulatory Role of RNA Structural Arrangement in Alternative Splicing Control. *Int. J. Mol. Sci.* **2020**, *21*, 5161. [CrossRef] [PubMed]
100. Mertes, C.; Scheller, I.F.; Yépez, V.A.; Çelik, M.H.; Liang, Y.; Kremer, L.S.; Gusic, M.; Prokisch, H.; Gagneur, J. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* **2021**, *12*, 1–13. [CrossRef] [PubMed]
101. Hu, Q.; Greene, C.S. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pac. Symp. Biocomput.* **2019**, *24*, 362–373. [CrossRef]