GigaScience

Genomic data imputation with variational autoencoders

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00059
Full Title:	Genomic data imputation with variational autoencoders
Article Type:	Technical Note
Funding Information:	National Institute of Biomedical Imaging and Bioengineering (R01 EB020527) — Mr. Olivier Gevaert National Institute of Biomedical Imaging and Bioengineering (R56 EB020527) — Mr. Olivier Gevaert National Cancer Institute (US) (U01 CA217851) — Mr. Olivier Gevaert National Cancer Institute (U01 CA199241) — Mr. Olivier Gevaert
Abstract:	As missing values are frequently present in genomic data, practical methods to handle missing data are necessary for downstream analyses that require complete datasets. State-of-the-art imputation techniques including Singular Value Decomposition (SVD) and K-Nearest Neighbors (KNN) based methods can be computationally expensive for large datasets and it is difficult to modify these algorithms to handle certain missing-not-at-random cases. In this work, we use a deep learning framework based on the variational autoencoder (VAE) for genomic missing value imputation and demonstrate its effectiveness in transcriptome and methylome data analysis. We show that in multiple simulated missing scenarios, VAE achieves similar or better performances than the most widely used imputation standards, while having computational advantage at evaluation time. When dealing with missing-not-at-random, e.g. low values are missing, we develop simple yet effective methodologies to leverage the prior knowledge about missing data. Furthermore, we investigate the effect of varying latent space regularization strength in VAE on the imputation performances, and in this context show why VAE has a better imputation capacity compared to a regular deterministic autoencoder (AE).
Corresponding Author:	Olivier Gevaert UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Yeping Lina Qiu
First Author Secondary Information:	
Order of Authors:	Yeping Lina Qiu Hong Zheng Olivier Gevaert
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No

Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Genomic data imputation with variational autoencoders

Yeping Lina Qiu[1,2], Hong Zheng[1], Olivier Gevaert[1,3,*]

[1] Medicine – Center for Biomedical Informatics Research, Stanford University, Stanford, CA USA

[2] Department of Electrical Engineering, Stanford University, Stanford, CA, USA

[3] Biomedical Data Science, Stanford University, Stanford, CA, USA

* To whom correspondence should be addressed: ogevaert@stanford.edu

Abstract

As missing values are frequently present in genomic data, practical methods to handle missing data are necessary for downstream analyses that require complete datasets. State-of-the-art imputation techniques including Singular Value Decomposition (SVD) and K-Nearest Neighbors (KNN) based methods can be computationally expensive for large datasets and it is difficult to modify these algorithms to handle certain missing-not-at-random cases. In this work, we use a deep learning framework based on the variational autoencoder (VAE) for genomic missing value imputation and demonstrate its effectiveness in transcriptome and methylome data analysis. We show that in multiple simulated missing scenarios, VAE achieves similar or better performances than the most widely used imputation standards, while having computational advantage at evaluation time. When dealing with missing-not-at-random, e.g. low values are missing, we develop simple yet effective methodologies to leverage the prior knowledge about missing data. Furthermore, we investigate the effect of varying latent space regularization strength in VAE on the imputation performances, and in this context show why VAE has a better imputation capacity compared to a regular deterministic autoencoder (AE).

Introduction

The massive and diverse datasets in genomics have provided researches a rich resource to study the molecular basis of diseases. The profiling of gene expression and DNA methylation have enabled the identification of cancer driver genes or biomarkers (Byron, et al., 2016; Gevaert, et al., 2015; Kulis and Esteller, 2010; Litovkin, et al., 2015; Tomczak, et al., 2015; Zheng, et al., 2019). Many such studies on cancer genomics require complete datasets (Champion, et al., 2018). However, missing values are frequently present in these data due to various reasons including low resolution, missing probes, and artifacts (Baghfalaki, et al., 2016; Libbrecht and Noble, 2015). Therefore, practical methods to handle missing data in genomic datasets are needed for effective downstream analyses.

One way to complete the data matrices is to ignore missing values by removing` the entire feature if any of the samples has a missing value in that feature, but this is usually not a good strategy as the feature may contain useful information for other samples. The most preferable way to handle missing data is to impute their values in the pre-processing step. Many approaches have been proposed for this purpose (Moorthy, et al., 2019), including replacement using average values, estimation using weighted K-nearest neighbor (KNN) method (Faisal and Tutz, 2017; Troyanskaya, et al., 2001), and estimation using singular value decomposition (SVD) based methods (Troyanskaya, et al., 2001). KNN and SVD are two techniques that have been commonly used as benchmarks against new developments (Smaragdis, et al., 2011; Yu, et al., 2010). KNN imputes missing value of a feature in a given sample with the weighted average of the feature values in a number of similar samples, as calculated by some distance measure. SVD attempts to estimate data structure from the entire input including the samples with missing values, and fill in the missing values iteratively according to the global structure. For this reason, SVD is inefficient on large matrices in practice since new decompositions have to be estimated for each missing sample, which is a very time-consuming process. However, SVD serves as an important benchmarking method to determine how well other, faster methods perform compared to SVD.

In recent years, a branch of machine learning which emerged based on big data and deep artificial neural network architectures, usually referred to as deep learning, has advanced rapidly and shown great potential for applications in bioinformatics (Min, et al., 2017). Deep learning has been applied in areas including genomics studies (Arisdakessian, et al., 2019; Chen, et al., 2016; Leung, et al., 2014), biomedical imaging (Chen, et al., 2016), and biomedical signal processing (Wulsin, et al., 2011). Autoencoders (AE) are a deep learning based model which form the basis of various frameworks for missing value imputation, and they have shown promising results for genomic data, imaging data and industrial data applications (Beaulieu-Jones and Moore, 2017; Eraslan, et al., 2019; Jaques, et al., 2018; Mattei and Frellsen, 2019; McCoy, et al., 2018; Vincent, et al., 2008). However, a simple AE without regularization is rarely ranked among the competitors for data imputation (Costa, et al., 2018; Garciarena and Santana, 2017). When a simple AE only focuses on creating output close to the input without any constraints, the model may overfit on the training data instead of learning the latent structure, such as dependencies and regularities characteristic of the data distribution (Vincent, et al., 2008), which makes it unlikely to impute well given new samples. Denoising autoencoder

(DAE) is a type of autoencoder that specifically uses noise corruption to the input to create robust latent features (Vincent, et al., 2008). DAE has been extensively used in the application of data imputation (Beaulieu-Jones and Moore, 2017; Costa, et al., 2018). The corrupting noise introduced in the DAE can be in many different forms, such as masking noise, Gaussian noise, and salt-and-pepper noise (Vincent, et al., 2010).

Variational autoencoders (VAE) are a probabilistic autoencoder that has wide applications in image and text generation (Hu, et al., 2017; Kingma and Welling, 2013; Yeh, et al., 2017). VAE learns the distributions of latent space variables that make the model generate output similar to the input. VAE has primarily been used as a powerful generative tool, having the ability to produce realistic fake contents in images, sound signal or texts, that highly resemble the real life contents that they learn from. The generative power is made possible by regularizing the latent space (Kingma and Welling, 2013). Constraining the latent space distributions to be close to a standard Gaussian helps to achieve a smooth latent space where two close points in the latent space should lead to similar reconstructions, and any point sampled from the latent space should give a meaningful reconstruction (Ghosh, et al., 2019). VAE has been applied in genomic contexts such as latent space learning of gene expression data (Way and Greene, 2017). In addition, recent works have applied VAE on single cell RNA sequencing data for clustering, batch correction and differential expression analysis (Grønbech, et al., 2018; Lopez, et al., 2018). However, VAE has not been extensively studied for genomic data imputation for bulk RNA expression and DNA methylation data, while large amounts of retrospective genomic and epigenomic data are available through databases like the Gene Expression Omnibus (GEO) (Barrett, et al., 2012) and the Short Read Archive (SRA)(Wheeler, et al., 2006).

Here, we examine the VAE mechanism and its application to genomic missing value imputation with bulk transcriptome and methylome data. We show that for both missing-completely-at-random and missing-not-at-random cases in transcriptome data and methylome data, VAE achieves similar or better performances than the de facto standards, and thus is a strong alternative to traditional methods for data imputation (Aghdam, et al., 2017). We demonstrate that in a missing-not-at-random scenario which involves a commonly encountered covariate shift, a shift correction method can be implemented to improve VAE's extrapolation performance. Furthermore, we investigate the effect of latent space regularization on imputation with a modification of the variational autoencoder - β-VAE (Higgins, et al., 2017). In the context of β-VAE results, we provide insights on why VAE can achieve good imputation performance compared to a regular deterministic AE.

Materials and Methods

Datasets

We use two datasets to perform data imputation: pan-cancer RNA sequencing data from The Cancer Genome Atlas (TCGA) datasets (Malta, et al., 2018; Tomczak, et al., 2015), and DNA methylation data (Campbell, et al., 2018; Gevaert, et al., 2015; Stunnenberg, et al., 2016). Both datasets contain only numeric values. The RNA sequencing data is expressed in RPKM (Reads

Per Kilobase of transcript, per Million mapped reads), which is a normalized unit of transcript expression. The DNA methylation data contains the numeric values of the methylation level at each CpG site. In the pre-processing stage of the RNA sequencing data, we remove the genes with NA values and then normalize the data by log transformation and z-score transformation. The resulting feature dimension is 17176 genes. We use 667 glioma patient samples, including glioblastoma (GBM) and low-grade-glioma (LGG), for training and testing the missing value imputation framework. In pre-processing the DNA methylation data, we remove the NA values, and normalize the data by negative log transformation and z-score transformation. We use the smallest chromosome subset (Chromosome 22) so that the resulting data dimension is not prohibitive for benchmarking different computation methods. The resulting data has 21220 CpG sites and 206 samples.

Missing data simulations

Each dataset is split into 80%-20% for training and testing the imputation framework. The sample split for the RNA sequencing dataset is stratified by the disease type, and the split is random for the DNA methylation data since the samples are homogenous. The training data is a complete dataset without missing values. Missing values are introduced to the testing data in two forms: missing-completely-at-random (MCAR) and missing-not-at-random (MNAR) (Little and Rubin, 2019).

In the MCAR case, we randomly mask a number of elements in each row by replacing the original values with NAs. To test a range of missing severity, we make the number of masked elements amount to 5%, 10%, and 30% of the total number of elements respectively.

For the gene expression data, we simulate three MNAR scenarios, each of which has 5% of the total data values missing. In the first scenario, the masked values are concentrated at certain genes. Such genes are selected based on their GC content, which is the percentage of nitrogenous bases on a RNA fragment that are either guanine (G) or cytosine (C). Too high or too low GC content influences RNA sequencing coverage, and potentially results in missing values from these genes (Chen, et al., 2013). We select genes with GC-content at the highest 10% and randomly mask half of these values. In the second simulation case, certain genes are masked entirely. We randomly select 5% of the genes and mask all values from these genes in the testing data, and as a result, the corrupted data miss all values for specific genes. The third scenario is based on gene expression level (Conesa, et al., 2016). If a gene is expressed at a low level, the few reads generated from this gene may not be captured when the sequencing depth is low. Therefore, we consider a possible scenario where lowly expressed genes are prone to be missing. In the testing data we first choose gene expression values at the lowest 10% quantile, and then randomly mask half of these values.

For the DNA methylation data, we simulate two MNAR scenarios. The first scenario is complete missing of certain genes, which is the same as the case in gene expression data. In the second case, we mask CpG sites that have fewer coverage than a certain threshold. If the coverage is low, it suggests that we do not have enough sequencing depth to confidently measure the methylation percentage, and therefore the values are prone to be missing. We set the coverage threshold to six in our experiments.

For each simulation scenario described above, we create ten random trials to measure the average imputation performance. The uncorrupted testing data is used to compute the imputation RMSE.

Variational autoencoder

An autoencoder is an unsupervised deep neural network that is trained to reconstruct an input X by learning a function $h_{w,b}(X) \approx X$. This is done by minimizing the loss function between the input X and the network's output X': $L(X, X')$. The most common loss function is the root mean squared error:

$$L(X, X') = \sqrt{||X - X'||^2} \qquad (1)$$

An autoencoder consists of an encoder and a decoder. The encoder transforms the input to a latent representation, often such that the latent representation is in a much smaller dimension than the input (Ballard, 1987). The decoder then maps the latent embedding to the reconstruction of X. An autoencoder is often used as a dimensional reduction technique to learn useful representations of data (Sakurada and Yairi, 2014).

While a regular autoencoder learns a latent space representation of data, a variational autoencoder (VAE) learns a distribution in the latent space. VAE is often used as a generative model by sampling from the learnt latent space distribution and generating new samples that are similar in nature as the original data (Kingma and Welling, 2013). The assumption of VAE is that the distribution of data X, $P(X)$ is related to the distribution of the latent variable z, $P(z)$ by

$$P_\theta(X) = \int P_\theta(X|z)P(z)dz \qquad (2)$$

Here $P_\theta(X)$, also known as the marginal likelihood, is the probability of each data point in X under the entire generative process, parametrized by θ. The model aims to maximize $P_\theta(X)$ by optimizing the parameter θ so as to approximate the true distribution of data. In practice, $P_\theta(X|z)$ will be nearly zero for most z, and it is therefore more practical to learn a distribution $Q_\phi(z|X)$ which gives rise to z that is likely to produce X and then compute $P(X)$ from $E_{z\sim Q_\phi}P(X|z)$. P_θ (X) and $E_{z\sim Q_\phi}P(X|z)$ can be shown to have the following relationship (Kingma and Welling, 2013):

$$log\, P_\theta(X) - D[Q_\phi(z|X)||P_\theta(z|X)] = E_{z\sim Q_\phi}[logP_\theta(X|z)] - D[Q_\phi(z|X)||P(z)] \qquad (3)$$

The left hand side of equation (3) is the quantity we want to maximize, $\log P_\theta(X)$, plus an error term, which is the Kullback-Liebler divergence between the approximated posterior distribution $Q_\phi(z|X)$ and the true posterior distribution $P_\theta(z|X)$. The KL divergence is a measure of how one distribution is different from another one, and is always non-negative. Thus, maximizing the log likelihood $\log P(X)$ can be achieved by maximizing the evidence lower bound (ELBO):

$$ELBO = \log P_\theta(X) - \mathcal{D}[Q_\phi(z|X)||P_\theta(z|X)] \quad (4)$$

The right hand side of equation (3) is something we can optimize by a gradient descent algorithm. $P_\theta(X|z)$ is modeled by the decoder network of the VAE parametrized by θ, and $Q_\phi(z|X)$ is modeled by the encoder network parametrized by ϕ. For continuous value inputs, $P_\theta(X|z)$ and $Q_\phi(z|X)$ are most commonly assumed to be Gaussian distributions (Ghosh, et al., 2019). $P(z)$ is a fixed prior distribution and assumed to be a standard multivariate normal distribution $\mathcal{N}(0, I)$. The first term $E_{z\sim Q_\phi}[logP_\theta(X|z)]$ is the expectation of the log probability of X given the encoder's output. Maximizing this term is equivalent to minimizing the reconstruction error of the autoencoder. The second term $D[Q_\phi(z|X)||P(z)]$ is the divergence between the approximated posterior distribution $Q_\phi(z|X)$ and the prior $P(z)$, and minimizing this term can be considered as adding a regularization term to prevent overfitting.

VAE is trained with the training data which follows a standard Gaussian distribution after z-score transformation. We impute missing values in the testing data with a trained VAE by an iterative process. Initially, the missing values are replaced with random values. Then the following sequence of steps are repeated until the iteration threshold is reached: compute the latent variable z distribution given input X with the encoder; take the mean of latent variable distribution as the input to the decoder, and compute the distribution of reconstructed data $\hat{X}$; take the mean of the reconstructed data distribution as the reconstructed values; replace the missing values with reconstructed values and leave non-missing values unchanged. The testing data is scaled by the training data mean and variance before the imputation iterations, and inverse scaled after imputation.

Variational autoencoder imputation with shift correction

Regular implementation of VAE has an underlying assumption that the training data follows the same distribution of testing data. Below, we will discuss how to modify this assumption to better impute missing-not-at-random scenarios.

Since the VAE learns the data distribution form the training data, the output of imputation also follows the learnt distribution, which is similar to the training data. When the missing values are drawn from a different distribution than the training data, the imputation performance will drop due to the distribution shift. In the missing-not-at-random simulations where half of the lowest 10% values are masked, the missing values are considered to be shifted from the original training data to a smaller mean.

The lowest value missing scenario, which is a common type of missing values in biomedical data, requires shift correction with VAE. Since the nature of the shift is relatively simple and known in advance, we leverage this knowledge to correct the shifting. Recall that in Equation (3), the underlying assumption is that the training data follows a Gaussian distribution X ~ $\mathcal{N}(\mu, \sigma)$, where μ and σ are the outputs of the decoder network which represent the mean and variance of the observed training data, as well as the missing data. When the lowest values are missing, the learnt distribution has larger mean than the actual missing data, causing the reconstructed $\hat{X}$ to have larger values. To correct this, we modify the assumption of training data

distribution to follow $\mathcal{N}(\mu + \lambda\sigma, \sigma)$ where μ and σ are the outputs of the decoder network which represent the mean and variance of the missing data, and λ is a hyperparameter. The mean of the observed training data is then shifted to $\mu + \lambda\sigma$.

To test the lowest 10% missing case, hyperparameter λ is selected on a validation data which simulates the lowest 10% missing case. In reality, we may not know the actual threshold and amount of missing in the testing data and thus cannot simulate the situation on the validation data precisely. For a various range of missing scenarios, where half of the lowest 5%, 10%, 20%, and 30% values are missing respectively, we impute with a single λ which is selected based on the lowest 10% missing case. We thereby determine if it is possible to select λ without precise knowledge of the missing scenario on the testing data.

β-VAE

β-VAE is a modification of the variational autoencoder with a focus to discover interpretable factorized latent factors (Higgins, et al., 2017). A hyperparameter beta is introduced to the VAE loss to balance the reconstruction loss term with the regularization loss term. The loss of β-VAE is defined as:

$$L_{\beta-\text{VAE}} = -E_{z\sim Q_\phi}[logP_\theta(X|z)] + \beta\mathcal{D}[Q_\phi(z|X)||P(z)] \qquad (5)$$

where β is a hyperparameter. β has been shown to affect certain image generation tasks (Burgess, et al., 2018). However, no prior work has investigated the effect of β on the imputation performance.

When β is 1, it is the same as VAE. When β>1, a stronger regularization is enforced, and the resulting latent space is smoother and more disentangled, which is a preferred property in certain learning tasks because more disentangled latent space has greater encoding efficiency (Higgins, et al., 2017).

On the other hand, when β=0, the regularization term is effectively removed. With the regularization term removed, the loss function only consists of the reconstruction loss term:

$$L_{\text{VAE}'} = -E_{z\sim Q_\phi}[logP_\theta(X|z)] \qquad (6)$$

which resembles the reconstruction loss function of a simple autoencoder (AE) without any regularization, that can usually be expressed in the mean squared error between the input X and the reconstruction X' (Kramer, 1991) :

$$L(X, X') = \|X - X'\|_2^2 \qquad (7)$$

However, the loss of VAE without the regularization term as shown in equation (6) has a key difference from the loss of a simple autoencoder shown in equation (7). If (6) is viewed from a deterministic perspective, it is easy to distinguish the difference.

With the assumption that P_θ and $Q_\phi$are Gaussian distributions,

$$P_\theta\,(X|z) \sim N\left(X\middle|\mu_\theta(z),\ diag\big(\sigma_\theta(z)\big)\right),$$

$$Q_\phi\,(z|X) \sim N\,(z|\mu_\phi(X),\ diag\left(\sigma_\phi(X)\right))$$

the loss in (6) can be computed as the mean squared error between inputs and their mean reconstructions output by the decoder (Ghosh, et al., 2019):

$$L_{\mathrm{VAE}'} = \|X - \mu_\theta(z)\|_2^2 \qquad (8)$$

Unlike the deterministic reconstruction X' in equation (7), z in equation (8) is stochastic. However, the stochasticity of z can be relegated to a random variable that does not depend on ϕ, so that we can view (8) from a deterministic perspective. Using the reparameterization trick (Kingma and Welling, 2013), z can be represented by:

$$z = \mu_\phi(X) + \sigma_\phi(X) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I) \qquad (9)$$

where $\odot$ is the element-wise product. Therefore the input to the decoder can be considered as the output of encoder $\mu_\phi(X)$ corrupted by a random gaussian noise ε multiplied by $\sigma_\phi(X)$. Consequently, the loss in (8) can be considered as the loss of a deterministic autoencoder, which but has noise injected to the latent space. In contrast, noise is not present in the deterministic regular AE loss in (7).

We perform three random missing experiments (5%, 10%, 30% missing) with β-VAE and vary the hyperparameter β from 0, 1, 4, to 10, to evaluate how β affects imputation accuracies. This will help us understand the VAE mechanism and how to use it in imputation.

Evaluation methods

To evaluate the VAE imputation framework, we compare it to other most commonly used missing value estimation methods: a K-nearest neighbor (KNN) method, and an iterative singular value decomposition (SVD) based method. KNN selects K number of samples which are most similar to the target sample with a missing gene based on Euclidean distance, and which all have values present in that gene. Imputation is a weighted average of the values of that gene in those K samples. We chose K=10 in our evaluations based on a study which reported that K in the range of 10-25 gave the best imputation results (Troyanskaya, et al., 2001). Next, the SVD method decomposes the data matrix to a linear combination of eigengenes and corresponding coefficients. Genes are regressed against L most significant eigengenes, during which process the missing genes are not used (Hastie, et al., 1999). The obtained coefficients are linearly multiplied by eigengenes to get a reconstruction with missing genes filled. This process is repeated until the total change in the matrix reaches a certain threshold. The reconstruction performance of SVD depends on the number of eigengenes selected for regression. We test a range of values and determine that the optimal performance is reached by full rank reconstruction. Hence we use full rank SVD in our evaluations.

We evaluate the root mean square error (RMSE) of the imputed data and uncorrupted ground truth,

$$RMSE = \frac{\sum_{i=1}^{n_{missing}} \sqrt{(x_i - x_i')^2}}{n_{missing}}$$

Where x_i is the ground truth of the masked value, and x_i' is the reconstructed value for the masked value.

To further evaluate the imputation effect on biomedical analysis, we compare the univariate correlation to two clinical variables imputed by different methods. We choose two variables: histology grade and survival outcome, which are available in the TCGA clinical data. We calculate the Spearman correlation coefficient between each gene and the histology grade variable, and the univariate cox regression coefficient of each gene with respect to the survival outcome. Then a concordance index is computed between the coefficient obtained from the imputed data by each method and the coefficients obtained from the ground truth. A higher concordance index indicates better resemblance to the true data.

Results

RMSE of imputation on RNA sequencing data

First, we evaluate the missing-completely-at-random cases at varying percentage 5%, 10%, and 30% random elements in the testing data were masked respectively, and models were compared on the reconstruction RMSE. In all tested random missing scenarios, VAE achieves better RMSE than KNN, and reaches similar or better performances than SVD (Figure 1a).

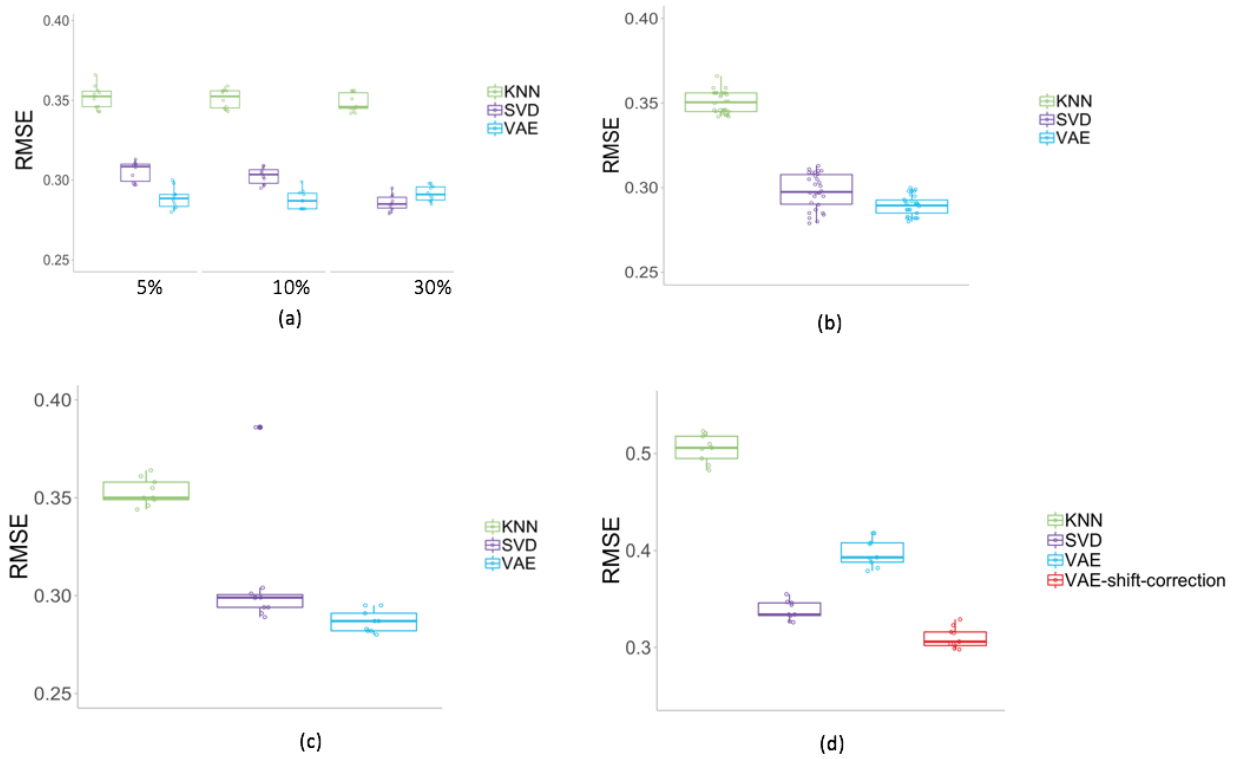


Figure 1 Imputation RMSE on the gene expression data for (a) missing-completely-at-random cases of 5%, 10% and 30%; (b) half of the highest 10% GC content genes missing case; (c) 5% genes entirely missing case; (d) half of the lowest 10% values missing case.

In the first missing-not-at-random simulation case, the masked values are confined to certain genes which have the highest 10% GC content. Random half of the genes whose GC content are in the top 10% miss their values in the testing data. VAE shows better reconstruction RMSE than KNN, and also achieves a slight advantage over SVD (Figure 1b). In the second case, 5% genes are masked entirely in the testing data. VAE again shows the best performance among competing methods (Figure 1c).

The final missing-no-at-random case is based on the gene expression values. The extreme values (either the lowest or the highest expression values) are masked from the testing data. As a result, the observed values in the testing data shifts its distribution from the training data, and results in a decreased performance of imputation. However, with shift-correction implementation, VAE again achieves similar or better imputation accuracy than other methods (Figure 1d).

The shift correction is robust to a range of low percentage missing scenarios

We further investigate the robustness of the shift correction parameter against a range of missing percentage on the lowest values. The shift correction parameter is selected based on a 10% lowest value missing scenario simulated on the validation data. We use the same selected

parameter to test on a range of missing scenarios, where half of the lowest 5%, 10%, 20%, and 30% values are missing respectively. All methods show worse prediction errors for smaller thresholds of missing values, because smaller thresholds indicate that the missing values are concentrated to smaller values, leading to a larger shifts in data distribution. We show that in these tested scenarios the shift correction VAE consistently achieves better result than KNN and SVD with the same λ (Figure 2). Therefore, λ selection does not need to exactly match the actual missing percentage, which is an advantage in real world implementations.

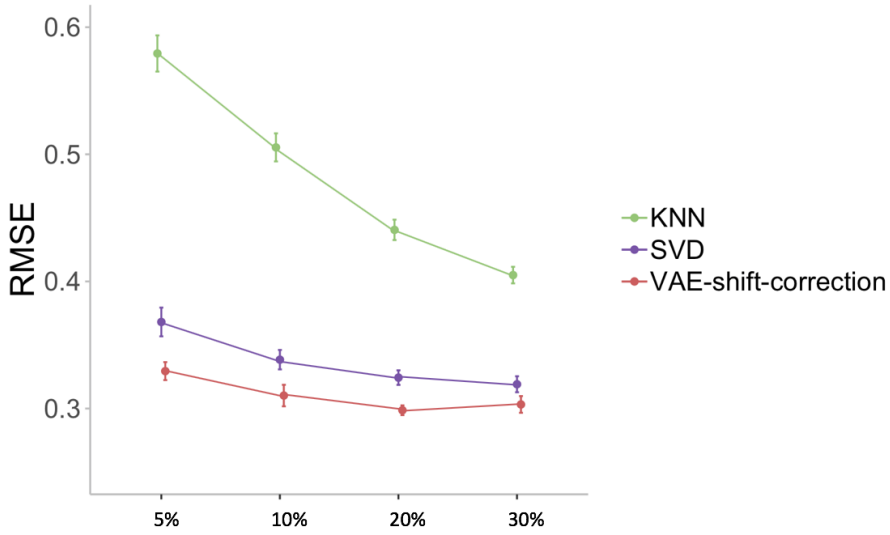


Figure 2. RMSE with 95% confidence interval for simulations where half of the lowest 5%, 10%, 20%, and 30% values are missing respectively. VAE-shift-correction results are achieved using a single λ which is selected based on the lowest 10% missing case.

RMSE of imputation on DNA methylation data

For the imputation on DNA methylation data, the comparative performance of the KNN, SVD and VAE methods shows similar performance compared to the gene expression data. In addition, these three methods show better performance than imputing with column mean. For missing-comeletely-at-random and block missing cases, VAE has similar performance as SVD, followed by KNN (Figure 3a, 3b). For the low coverage missing case, VAE achieves better RMSE than SVD and KNN (Figure 3c).

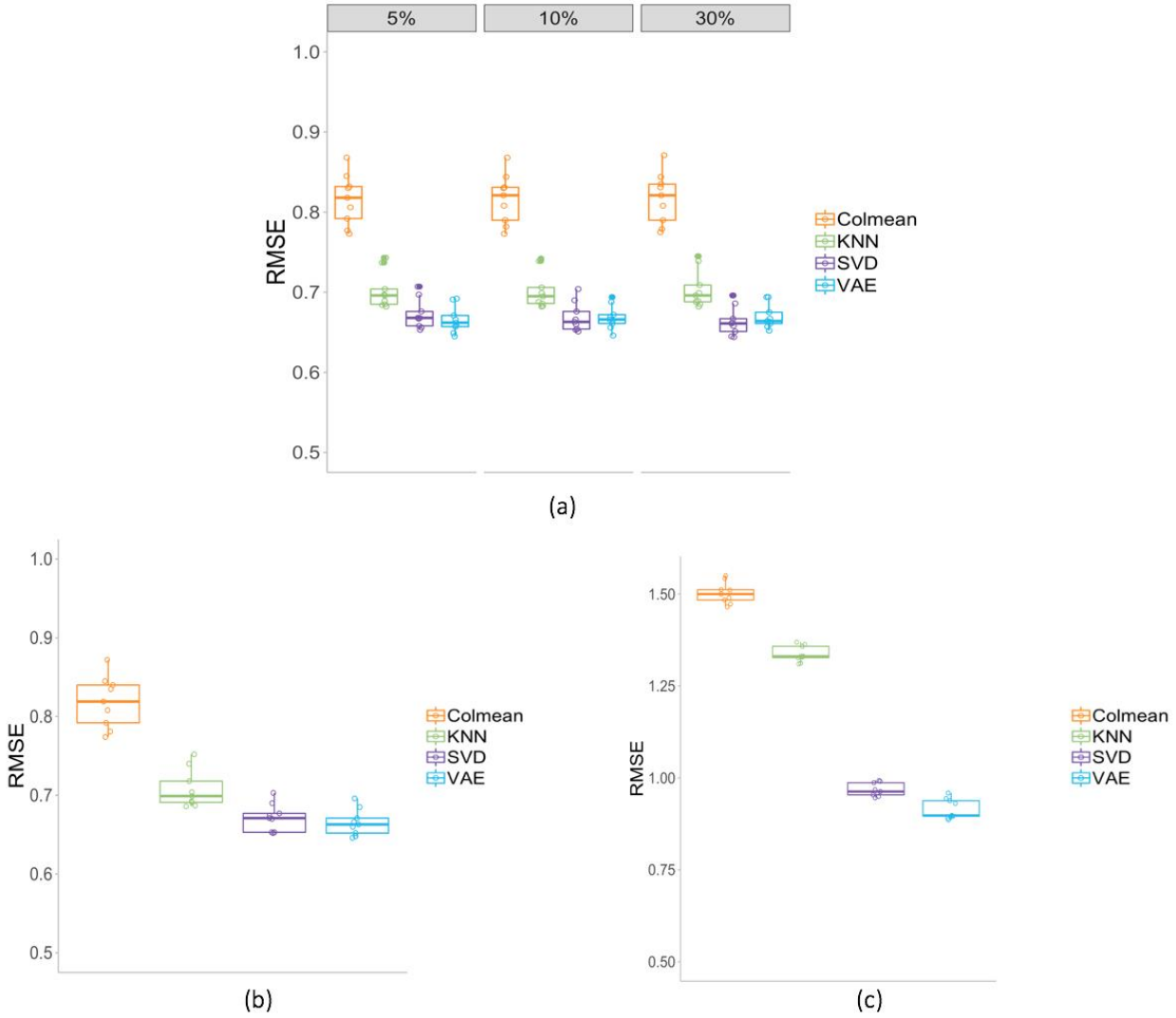


Figure 3. Imputation RMSE on the DNA methylation data for (a) missing-completely-at-random cases of 5%, 10% and 30%; (b) 5% genes entirely missing; (d) half of the coverage <6 CpG sites missing.

Correlation with clinical phenotypes

We investigate how closely the imputed data resembles the true data in terms of univariate correlation with respect to clinical variables including histology and survival outcome. Table 1 shows the concordance index between the correlation coefficient obtained from the imputed data by each method and the coefficients obtained from the ground truth. VAE and SVD achieve better concordance index than KNN, which likely indicates a better resemblance to true data in the context of biomedical analysis for molecular biologists interested in specific genes in the presence of missing values. Figure 4 illustrates pairwise difference between the coefficients obtained from the ground truth and the coefficients obtained from the imputed data by KNN and VAE respectively, and shows sharper peaks around zero for VAE in all cases for histology and in most cases for survival.

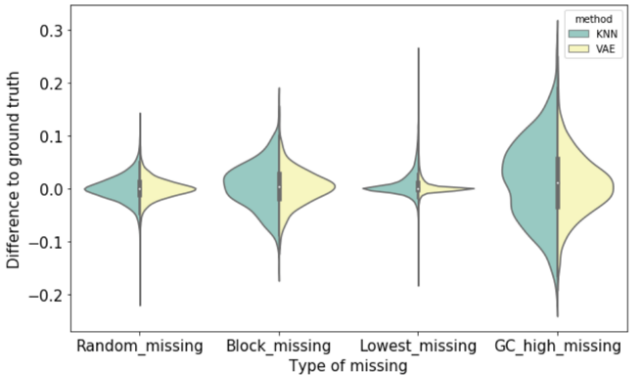
Table 1. Correlation with clinical phenotypes**:** (a) Pearson correlation coefficient with tumor histology grade; (b) Cox regression coefficient with survival outcome

(a)

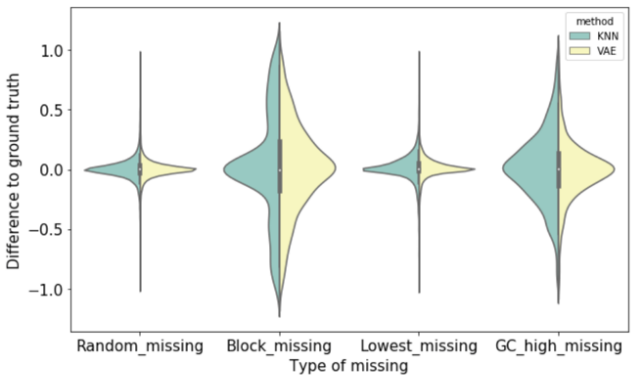
	KNN	SVD	VAE
Random missing	0.98	0.98	0.98
Highest GC content missing	0.95	**0.96**	**0.96**
Entire genes missing	0.92	**0.94**	0.93
Lowest value missing	0.98	**0.99**	**0.99**

(b)

	KNN	SVD	VAE
Random missing	0.97	0.97	0.97
Highest GC content missing	0.92	**0.93**	**0.93**
Entire genes missing	0.86	**0.92**	0.89
Lowest value missing	0.96	**0.97**	**0.97**



(a)



(b)

Figure 4. Pairwise difference between the coefficients obtained from the ground truth and the coefficients obtained from the imputed data by KNN and VAE: (a) Pearson correlation coefficients with histology grade; (b) Regression coefficients with survival outcome

Imputation time for new samples

The computation time for SVD or KNN to impute a single sample scales linearly with the dimension of the entire data matrix, while on the other hand, a VAE model can be pre-trained and applied directly to any given new sample to impute missing values. Once a VAE model is trained, the time to impute a new sample is almost negligible. VAE thus has the benefit of reducing computational cost especially at evaluation time.

Benchmark experiments are done on a 20 core cluster with Intel Xeon 2.40GHz CPUs, where the three methods are used to impute 100 samples in a gene expression matrix that consists of 6600 samples and 17176 genes. It takes an average of 2800 seconds to train the VAE network. In terms of evaluation time, the KNN method takes on average 8400 seconds, while SVD takes 36900 seconds, and VAE takes only 60 seconds, showing that VAE is several orders of magnitude faster at evaluation time.

β-VAE and deterministic autoencoder

We perform three random missing experiments with β-VAE and vary the hyperparameter β from 0, 1, 4, to 10. Figure 5 shows that imputation results are similar for β=0 and β=1, while increasing β to larger values worsens the prediction accuracies.

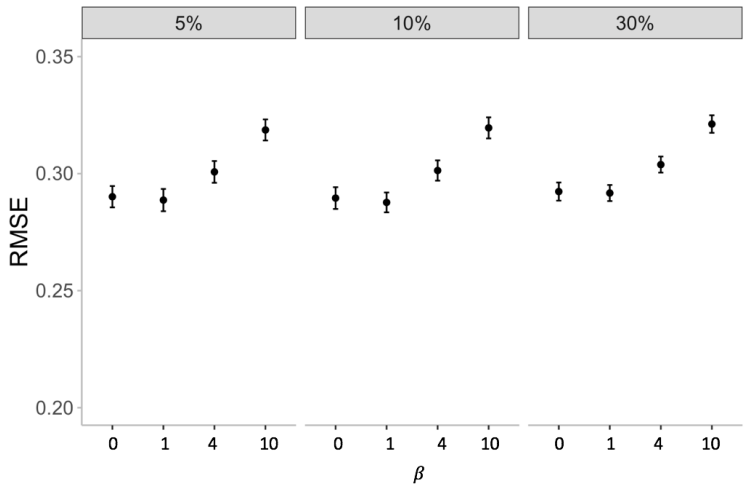


Figure 5. Imputation RMSE error of β-VAE on 5%, 10% and 30% random missing of gene expression, with β = 0, 1, 4, and 10, denoting increasing strength of regularization.

The fact that β >1 leads to worse imputation errors may be explained by considering the total loss of VAE (right hand side of equation (3)), consisting of the reconstruction loss and regularization loss, as a tradeoff between reconstruction quality and latent space coding efficiency. If a greater emphasis is put on latent space regularization, the reconstruction quality suffers.

When β=0, the model is a VAE without regularization. We observe that the imputation performance is similar to vanilla VAE ($\beta = 1$). Therefore, for the purpose of imputation, latent space regularization is not essential, and it is possible to remove this term in the VAE implementation.

From previous discussion, β-VAE with β=0 can be viewed as a deterministic autoencoder with noise injected to the latent space. We find that with a simple deterministic AE without noise injection, the imputation iterations cannot converge and the resulting RMSE is very large (not shown because non-convergence). This suggests that the noise injection to the latent space enables the imputation ability of the autoencoder.

Conclusion

We have described a deep learning imputation framework for transcriptome and methylome data using a variational autoencoder (VAE). We implement a shift correction method to improve VAE imputation performance on a commonly encountered missing-not-at-random scenario. We demonstrate that the proposed framework is competitive with SVD, which is a time-inefficient method for real world scenarios. We also show that VAE outperforms KNN in multiple scenarios such as transcriptome and methylome data. VAE thus can be an important tool to analyze the large amounts of publicly available data from 1000s of studies that are publicly available in the Gene Expression omnibus (Barrett, et al., 2012).

We show that noise addition to the latent space is the essential mechanism that enables VAE's good imputation performance, compared to a regular deterministic AE. The method of noise injection during training is reminiscent of denoising autoencoders (DAE). However, the noise addition for VAE and DAE are different. First, the noise in VAE depends on the input, whereas the DAE noise is independent of the input. Second, although noise addition to intermediate layers has been proposed in stacked denoising autoencoders for the purpose of representation learning (Vincent, et al., 2010), in most data imputation applications noise has only been added to the input layer of DAE (Costa, et al., 2018; Gondara and Wang, 2017). On the other hand, noise is added to the latent space layer in VAE. It is not in the scope of this paper to evaluate how different noise addition schemes impact imputation and compare their performances. However, this may be worth exploring in future work.

We also provide insight on the effect of latent space regularization on imputation performance. We show that increasing latent space regularization in the VAE implementation leads to larger error, and thus should be avoided in the imputation tasks. On the other hand, based on the hypothesis that there is a tradeoff between reconstruction quality and desired latent space property regulated by β, it can be expected that removing the regularization term (β=0) may even improve the vanilla VAE's (β=1) imputation performance. It is worth noting that such phenomenon did not occur. Future work is needed to fully understand the effect of β in the range between 0 and 1.

Finally, in the context of imputing large dataset with high dimensional features, VAE has the potential benefit of reducing computational cost at evaluation time compared to SVD and KNN. This is because an autoencoder model can be pre-trained and applied directly to new samples, while SVD and KNN require computing the entire matrix each time a new sample is given.

Availability of Data and Materials

All data used in this manuscript are held in a public repository. Gene expression data can be found with URL: https://www.synapse.org/#!Synapse:syn4976369.2. DNA methylation data can be found with URL: https://www.rnbeads.org/methylomes.html. Code is available at https://github.com/gevaertlab/BetaVAEImputation.

Declarations

Funding

Research reported in this publication was supported by the Fund for Innovation in Cancer Informatics (www.the-ici-fund.org), by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIBIB https://www.nibib.nih.gov/), R01 EB020527 and R56 EB020527, and the National Cancer Institute (NCI https://www.cancer.gov/), U01 CA217851 and U01 CA199241, all to OG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

Aghdam, R., *et al.* The Ability of Different Imputation Methods to Preserve the Significant Genes and Pathways in Cancer. *Genomics, Proteomics & Bioinformatics* 2017;15(6):396-404.

Arisdakessian, C., *et al.* DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome biology* 2019;20(1):1-14.

Baghfalaki, T., Ganjali, M. and Berridge, D. Missing Value Imputation for RNA-Sequencing Data Using Statistical Models:
A Comparative Study. *Journal of Statistical Theory and Applications* 2016;15.

Ballard, D.H. Modular learning in neural networks. In, *Proceedings of the sixth National conference on Artificial intelligence - Volume 1*. Seattle, Washington: AAAI Press; 1987. p. 279-284.

Barrett, T., *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* 2012;41(D1):D991-D995.

Beaulieu-Jones, B.K. and Moore, J.H. Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. *Pac Symp Biocomput* 2017;22:207-218.

Burgess, C.P., *et al.* Understanding disentangling in $\beta $-VAE. *arXiv preprint arXiv:1804.03599* 2018.

Byron, S.A., *et al.* Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 2016;17(5):257-271.

Campbell, J.D., *et al.* Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell reports* 2018;23(1):194-212. e196.

Champion, M., *et al.* Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. *EBioMedicine* 2018;27:156-166.

Chen, C.L., *et al.* Deep Learning in Label-free Cell Classification. *Scientific Reports* 2016;6:21471.

Chen, Y., *et al.* Gene expression inference with deep learning. *Bioinformatics* 2016;32(12):1832-1839.

Chen, Y.C., *et al.* Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 2013;8(4):e62856.

Conesa, A., *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17:13.

Costa, A.F., *et al.* Missing data imputation via denoising autoencoders: the untold story. In, *International Symposium on Intelligent Data Analysis*. Springer; 2018. p. 87-98.

Eraslan, G., *et al.* Single-cell RNA-seq denoising using a deep count autoencoder. *Nature communications* 2019;10(1):390.

Faisal, S. and Tutz, G. Missing value imputation for gene expression data by tailored nearest neighbors. *Stat Appl Genet Mol Biol* 2017;16(2):95-106.

Garciarena, U. and Santana, R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications* 2017;89:52-65.

Gevaert, O., Tibshirani, R. and Plevritis, S.K. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biol* 2015;16:17.

Gevaert, O., Tibshirani, R. and Plevritis, S.K. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome biology* 2015;16(1):17.
Ghosh, P., *et al.* From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436* 2019.
Gondara, L. and Wang, K. Multiple imputation using deep denoising autoencoders. *arXiv preprint arXiv:1705.02737* 2017.
Grønbech, C.H., *et al.* scVAE: Variational auto-encoders for single-cell gene expression data. *bioRxiv* 2018:318295.
Hastie, T., *et al.* Imputingmissing data for gene expression arrays. 1999.
Higgins, I., *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR* 2017;2(5):6.
Hu, Z., *et al.* Toward controlled generation of text. In, *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org; 2017. p. 1587-1596.
Jaques, N., *et al.* Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. 2018.
Kingma, D.P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* 2013.
Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 1991;37(2):233-243.
Kulis, M. and Esteller, M. DNA methylation and cancer. *Adv Genet* 2010;70:27-56.
Leung, M.K., *et al.* Deep learning of the tissue-regulated splicing code. *Bioinformatics* 2014;30(12):i121-129.
Libbrecht, M.W. and Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;16(6):321-332.
Litovkin, K., *et al.* DNA Methylation-Guided Prediction of Clinical Failure in High-Risk Prostate Cancer. *PLoS One* 2015;10(6):e0130651.
Little, R.J. and Rubin, D.B. Statistical analysis with missing data. John Wiley & Sons; 2019.
Lopez, R., *et al.* Deep generative modeling for single-cell transcriptomics. *Nature methods* 2018;15(12):1053.
Malta, T.M., *et al.* Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 2018;173(2):338-354. e315.
Mattei, P.-A. and Frellsen, J. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In, *International Conference on Machine Learning*. 2019. p. 4413-4423.
McCoy, J.T., Kroon, S. and Auret, L. Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit. *IFAC-PapersOnLine* 2018;51(21):141-146.
Min, S., Lee, B. and Yoon, S. Deep learning in bioinformatics. *Brief Bioinform* 2017;18(5):851-869.
Moorthy, K., *et al.* Missing-values imputation algorithms for microarray gene expression data. In, *Microarray Bioinformatics*. Springer; 2019. p. 255-266.
Sakurada, M. and Yairi, T. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In, *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. Gold Coast, Australia QLD, Australia: ACM; 2014. p. 4-11.
Smaragdis, P., Raj, B. and Shashanka, M. Missing Data Imputation for Time-Frequency Representations of Audio Signals. *Journal of Signal Processing Systems* 2011;65(3):361-370.
Stunnenberg, H.G., *et al.* The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* 2016;167(5):1145-1149.

Tomczak, K., Czerwinska, P. and Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;19(1A):A68-77.
Troyanskaya, O., *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;17(6):520-525.
Vincent, P., *et al.* Extracting and composing robust features with denoising autoencoders. In, *Proceedings of the 25th international conference on Machine learning*. Helsinki, Finland: ACM; 2008. p. 1096-1103.
Vincent, P., *et al.* Extracting and composing robust features with denoising autoencoders. In, *Proceedings of the 25th international conference on Machine learning*. ACM; 2008. p. 1096-1103.
Vincent, P., *et al.* Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* 2010;11:3371-3408.
Way, G.P. and Greene, C.S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *BioRxiv* 2017:174474.
Wheeler, D.L., *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* 2006;35(suppl_1):D5-D12.
Wulsin, D.F., *et al.* Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *J Neural Eng* 2011;8(3):036015.
Yeh, R.A., *et al.* Semantic image inpainting with deep generative models. In, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. p. 5485-5493.
Yu, T., Peng, H. and Sun, W. Incorporating nonlinear relationships in microarray missing value imputation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010;8(3):723-731.
Zheng, H., *et al.* Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience* 2019;8(12):giz145.