frontiers in Genetics

ORIGINAL RESEARCH
published: 13 April 2021
doi: 10.3389/fgene.2021.624128

Deep Learning Enables Fast and Accurate Imputation of Gene Expression

Ramon Viñas[1]*, *Tiago Azevedo*[1], *Eric R. Gamazon*[2,3,4]* and *Pietro Liò*[1]*

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [2] Vanderbilt Genetics Institute and Data Science Institute, VUMC, Nashville, TN, United States, [3] MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, [4] Clare Hall, University of Cambridge, Cambridge, United Kingdom

A question of fundamental biological significance is to what extent the expression of a subset of genes can be used to recover the full transcriptome, with important implications for biological discovery and clinical application. To address this challenge, we propose two novel deep learning methods, PMI and GAIN-GTEx, for gene expression imputation. In order to increase the applicability of our approach, we leverage data from GTEx v8, a reference resource that has generated a comprehensive collection of transcriptomes from a diverse set of human tissues. We show that our approaches compare favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime in two case studies and two imputation scenarios. In comparison conducted on the protein-coding genes, PMI attains the highest performance in inductive imputation whereas GAIN-GTEx outperforms the other methods in in-place imputation. Furthermore, our results indicate strong generalization on RNA-Seq data from 3 cancer types across varying levels of missingness. Our work can facilitate a cost-effective integration of large-scale RNA biorepositories into genomic studies of disease, with high applicability across diverse tissue types.

Keywords: gene expression, transcriptomics, imputation, generative adversarial networks, machine learning, RNA-seq, GTEx, deep learning

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Miguel Andrade,
Johannes Gutenberg University Mainz, Germany
Jidong Lang,
Geneis (Beijing) Co. Ltd, China

***Correspondence:**
Ramon Viñas
rv340@cam.ac.uk
Eric R. Gamazon
ericgamazon@gmail.com
Pietro Liò
pl219@cam.ac.uk

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *30 October 2020*
Accepted: *12 March 2021*
Published: *13 April 2021*

Citation:
Viñas R, Azevedo T, Gamazon ER and Liò P (2021) Deep Learning Enables Fast and Accurate Imputation of Gene Expression. Front. Genet. 12:624128. doi: 10.3389/fgene.2021.624128

1. INTRODUCTION

High-throughput profiling of the transcriptome has revolutionized discovery methods in the biological sciences. The resulting gene expression measurements can be used to uncover disease mechanisms (Emilsson et al., 2008; Cookson et al., 2009; Gamazon et al., 2018), propose novel drug targets (Evans and Relling, 2004; Sirota et al., 2011), provide a basis for comparative genomics (King and Wilson, 1975; Colbran et al., 2019), and motivate a wide range of fundamental biological problems. In parallel, methods that learn to represent the expression manifold can improve our mechanistic understanding of complex traits, with potential methodological and technological applications, including organs-on-chips (Low et al., 2020) and synthetic biology (Gupta and Zou, 2019), and the integration of heterogeneous transcriptomics datasets.

A question of fundamental biological significance is to what extent the expression of a subset of genes can be used to recover the full transcriptome with minimal reconstruction error. Genes that participate in similar biological processes or that have shared molecular function are likely to have similar expression profiles (Zhang and Horvath, 2005), prompting the question of gene expression prediction from a minimal subset of genes. Moreover, gene expression measurements may suffer

from unreliable values because some regions of the genome are extremely challenging to interrogate due to high genomic complexity or sequence homology (Conesa et al., 2016), further highlighting the need for accurate imputation. Moreover, most gene expression studies continue to be performed with specimens derived from peripheral blood or a convenient surrogate (e.g., lymphoblastoid cell lines; LCLs) due to the difficulty of collecting some tissues. However, gene expression may be tissue or cell-type specific, potentially limiting the utility of a proxy tissue.

The missing data problem can adversely affect downstream gene expression analysis. The simple approach of excluding samples with missing data from the analysis can lead to a substantial loss in statistical power. Dimensionality reduction approaches such as principal component analysis (PCA) and singular value decomposition (SVD) (Wall et al., 2003) cannot be applied to gene expression data with missing values. Clustering methods, a mainstay of genomics, such as *k*-means and hierarchical clustering may become unstable even with a few missing values (Troyanskaya et al., 2001).

To address these challenges, we develop two deep learning approaches to gene expression imputation. In both cases, we present an architecture that recovers missing expression data for multiple tissue types under different levels of missingness. In contrast to existing linear methods for deconfounding gene expression (Øystein Sørensen et al., 2018), our methods integrate covariates (global determinants of gene expression; Stegle et al., 2012) to account for their non-linear effects. In particular, a characteristic feature of our architectures is the use of word embeddings (Mikolov et al., 2013) to learn rich and distributed representations for the tissue types and other covariates. To enlarge the possibility and scale of a study's expression data (e.g., by including samples from highly inaccessible tissues), we train our model on RNA-Seq data from the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2015; GTEx Consortium, 2017), a reference resource (v8) that has generated a comprehensive collection of human transcriptome data in a diverse set of tissues.

We show that the proposed approaches compare favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime. In performance comparison on the protein-coding genes, GAIN-GTEx outperforms all the other methods in in-place imputation while PMI displays the highest performance in inductive imputation. Furthermore, we demonstrate that our methods are highly applicable across diverse tissues and varying levels of missingness. Finally, to analyse the cross-study relevance of our approach, we perform imputation on gene expression data from The Cancer Genome Atlas (TCGA; Weinstein et al., 2013) and show that our approach is robust when applied to independent RNA-Seq data.

2. METHODS

In this section, we introduce two deep learning approaches for gene expression imputation with broad applicability, allowing us to investigate their strengths and weaknesses in several realistic scenarios. Throughout the remainder of the paper, we use script letters to denote sets (e.g., $\mathcal{D}$), upper-case bold symbols to denote matrices or random variables (e.g., $\mathbf{X}$), and lower-case bold symbols to denote column vectors (e.g., $\mathbf{x}$ or $\bar{\mathbf{q}}_j$). The rest of the symbols (e.g., $\bar{q}_{jk}$, G, or f) denote scalar values or functions.

2.1. Problem Formulation

Consider a dataset $\mathcal{D} = \{(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$, where $\tilde{\mathbf{x}} \in \mathbb{R}^n$ represents a vector of gene expression values with missing components; $\mathbf{m} \in \{0,1\}^n$ is a mask indicating which components of the original vector of expression values $\mathbf{x}$ are missing or observed; n is the number of genes; and $\mathbf{q} \in \mathbb{N}^c$ and $\mathbf{r} \in \mathbb{R}^k$ are vectors of c categorical (e.g., tissue type or sex) and k quantitative covariates (e.g., age), respectively. Our goal is to recover the original gene expression vector $\mathbf{x} \in \mathbb{R}^n$ by modeling the conditional probability distribution $P(\mathbf{X} = \mathbf{x} | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \mathbf{M} = \mathbf{m}, \mathbf{R} = \mathbf{r}, \mathbf{Q} = \mathbf{q})$, where the upper-case symbols denote the corresponding random variables.

2.2. Pseudo-Mask Imputation

We first introduce a novel imputation method named Pseudo-Mask Imputer (PMI).

Formulation. Let $\tilde{\mathbf{x}} = \mathbf{m} \odot \mathbf{x} \in \mathbb{R}^n$ be a vector of gene expression values whose missing components are indicated by a mask vector $\mathbf{m} \in \{0,1\}^n$. Our model is a function $f : \mathbb{R}^n \times \{0,1\}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that imputes the missing expression values $(\mathbf{1} - \mathbf{m}) \odot \mathbf{x}$ as follows:

$$\bar{\mathbf{x}} = f(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q}). \tag{1}$$

Here $\odot$ denotes element-wise multiplication. The recovered vector of gene expression values is then given by $\mathbf{m} \odot \tilde{\mathbf{x}} + (\mathbf{1} - \mathbf{m}) \odot \bar{\mathbf{x}}$.

Optimization. We optimize the model to maximize the imputation performance on a dynamic subset of observed, *pseudo-missing* components. In particular, we first generate a *pseudo-mask* $\tilde{\mathbf{m}}$ as follows:

$$\tilde{\mathbf{m}} = \mathbf{m} \odot \mathbf{b} \qquad \mathbf{b} \sim B(1, p) \qquad p \sim U(\alpha, \beta), \tag{2}$$

where $\mathbf{b} \in \{0,1\}^n$ is a vector sampled from a Bernoulli distribution B and $\alpha \in [0,1]$ and $\beta \in [\alpha, 1]$ are hyperparameters that parameterize a uniform distribution U. Using the *pseudo-mask* $\tilde{\mathbf{m}}$, we split the observed expression values into a set of *pseudo-observed* components $\tilde{\mathbf{x}}$ and a set of *pseudo-missing* components $\tilde{\mathbf{y}}$:

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \tilde{\mathbf{m}} \qquad \tilde{\mathbf{y}} = \mathbf{x} \odot \mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}}), \tag{3}$$

The imputed components are then given by $\bar{\mathbf{x}} = f(\tilde{\mathbf{x}}, \tilde{\mathbf{m}}, \mathbf{r}, \mathbf{q})$. We optimize our model to minimize the mean squared error between the ground truth and the imputed *pseudo-missing* components:

$$\mathcal{L}(\bar{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{m}}) = \frac{1}{Z} \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)^\top (\bar{\mathbf{x}} - \tilde{\mathbf{y}})^2, \tag{4}$$

where $Z = \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)^\top \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)$ is a normalization term. We summarize our training algorithm in Algorithm 1.

Importantly, the *pseudo-mask* mechanism generates different sets of *pseudo-observed* components for each input example, effectively enlarging the number of training samples. Specifically, the hyperparameters α and β control the fraction of *pseudo-observed* and *pseudo-missing* components through the probability $p \sim U(\alpha, \beta)$. On one hand, a low probability p yields sparse *pseudo-observed* vectors $\hat{\mathbf{x}}$, resulting in fast convergence but high bias. On the other hand, a high probability p yields denser *pseudo-observed* vectors $\hat{\mathbf{x}}$, resulting in low bias but slower convergence. At inference time, p is set to 1 and the *pseudo-mask* $\tilde{\mathbf{m}}$ is equal to the input mask $\mathbf{m}$.

Algorithm 1: Training algorithm

Input: Input dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$, batch size B, hyperparameters α and β

- Initialise parameters of the model f

while *not convergence criteria reached* **do**

- Sample mini-batch: $\{(\mathbf{x}^{(i)}, \mathbf{m}^{(i)}, \mathbf{r}^{(i)}, \mathbf{q}^{(i)})\}_{i=1}^{B} \sim \mathcal{D}$
- Sample *pseudo-mask* for each example of the mini-batch: $p^{(i)} \sim U(\alpha, \beta)$ $\mathbf{b}^{(i)} \sim B(1, p^{(i)})$ $\tilde{\mathbf{m}}^{(i)} = \mathbf{m}^{(i)} \odot \mathbf{b}^{(i)}$
- Split components into *pseudo-observed* and *pseudo-missing*: $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} \odot \tilde{\mathbf{m}}^{(i)}$ $\tilde{\mathbf{y}}^{(i)} = \mathbf{x}^{(i)} \odot \mathbf{m}^{(i)} \odot (\mathbf{1} - \tilde{\mathbf{m}}^{(i)})$
- Impute *pseudo-missing* components: $\bar{\mathbf{x}}^{(i)} = f(\tilde{\mathbf{x}}^{(i)}, \tilde{\mathbf{m}}^{(i)}, \mathbf{r}^{(i)}, \mathbf{q}^{(i)})$
- Optimise the model by descending its stochastic gradient: $\nabla \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}(\bar{\mathbf{x}}^{(i)}, \tilde{\mathbf{y}}^{(i)}, \mathbf{m}^{(i)}, \tilde{\mathbf{m}}^{(i)})$

end

Architecture. We model the imputer f as a neural network. We first describe how we use word embeddings, a distinctive feature that allows learning rich, dense representations for the different tissue types and, more generally, for all the covariates $\mathbf{q} \in \mathbb{N}^c$.

Formally, let q_j be a categorical covariate (e.g., tissue type) with vocabulary size v_j, that is, $q_j \in \{1, 2, \ldots, v_j\}$, where each value in the vocabulary $\{1, 2, \ldots, v_j\}$ represents a different category (e.g., whole blood or kidney). Let $\bar{\mathbf{q}}_j \in \{0, 1\}^{v_j}$ be a one-hot vector such that $\bar{q}_{jk} = 1$ if $q_j = k$ and $\bar{q}_{jk} = 0$ otherwise. Let d_j be the dimensionality of the embeddings for covariate j. We obtain a vector of embeddings $\mathbf{e}_j \in \mathbb{R}^{d_j}$ as follows:

$$\mathbf{e}_j = \bar{\mathbf{q}}_j^{\top} \mathbf{W}_j, \tag{5}$$

where each $\mathbf{W}_j \in \mathbb{R}^{v_j \times d_j}$ is a matrix of learnable weights. Essentially, this operation describes a lookup search in a dictionary with v_j entries, where each entry contains a learnable d_j-dimensional vector of embeddings that characterize each of the possible values that q_j can take. To obtain a global collection of embeddings $\mathbf{e}$, we concatenate all the vectors $\mathbf{e}_j$ for each categorical covariate j:

$$\mathbf{e} = \Big\|_{j=1}^{c} \mathbf{e}_j, \tag{6}$$

where c is the number of categorical covariates and $\|$ represents the concatenation operator. We then use the learnable embeddings $\mathbf{e}$ in downstream tasks.

In terms of the architecture, we model f as follows:

$$f(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q}) = \mathrm{MLP}(\tilde{\mathbf{x}} \| \mathbf{m} \| \mathbf{r} \| \mathbf{e}), \tag{7}$$

where MLP denotes a multilayer perceptron and $\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbf{m} \in \mathbb{R}^n$ is the masked gene expression. **Figure 1** shows the architecture of the model.

2.3. Generative Adversarial Imputation Networks

The second method, which we call GAIN-GTEx, is based on Generative Adversarial Imputation Nets (GAIN; Yoon et al., 2018). Generative Adversarial Networks have previously been used to synthesize transcriptomics *in-silico* (Marouf et al., 2020; Viñas et al., 2021), but to our knowledge their applicability to gene expression imputation is yet to be studied. Similar to generative adversarial networks (GANs; Goodfellow et al., 2014), GAIN estimates a generative model via an adversarial process driven by the competition between two players, the *generator* and the *discriminator*.

Generator. The generator aims at recovering missing data from partial gene expression observations, producing samples from the conditional $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Formally, we define the generator as a function $G : \mathbb{R}^n \times \mathbb{R}^n \times \{0, 1\}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that imputes expression values as follows:

$$\bar{\mathbf{x}} = G(\mathbf{x} \odot \mathbf{m}, \mathbf{z} \odot (\mathbf{1} - \mathbf{m}), \mathbf{m}, \mathbf{r}, \mathbf{q}), \tag{8}$$

where $\mathbf{z} \in \mathbb{R}^n$ is a vector sampled from a fixed noise distribution. Similar to GAIN, we mask the n-dimensional noise vector as $\mathbf{z} \odot (\mathbf{1} - \mathbf{m})$, encouraging a bijective association between noise components and genes. Before passing the output $\bar{\mathbf{x}}$ to the discriminator, we replace the prediction for the non-missing components by the original, observed expression values:

$$\hat{\mathbf{x}} = \mathbf{m} \odot \tilde{\mathbf{x}} + (\mathbf{1} - \mathbf{m}) \odot \bar{\mathbf{x}}. \tag{9}$$

Discriminator. The discriminator takes the imputed samples $\hat{\mathbf{x}}$ and attempts to distinguish whether the expression value of each gene has been observed or produced by the generator. This is in contrast to the original GAN discriminator, which receives information from two input streams (generator and data distribution) and attempts to distinguish the true input source.

Formally, the discriminator is a function $D : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that outputs the probabilities $\hat{\mathbf{y}} \in \mathbb{R}^n$:

$$\hat{\mathbf{y}} = D(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{r}, \mathbf{q}), \tag{10}$$

where the i-th component $\hat{y}_i$ is the probability of gene i being observed (as opposed to being imputed by the generator) for

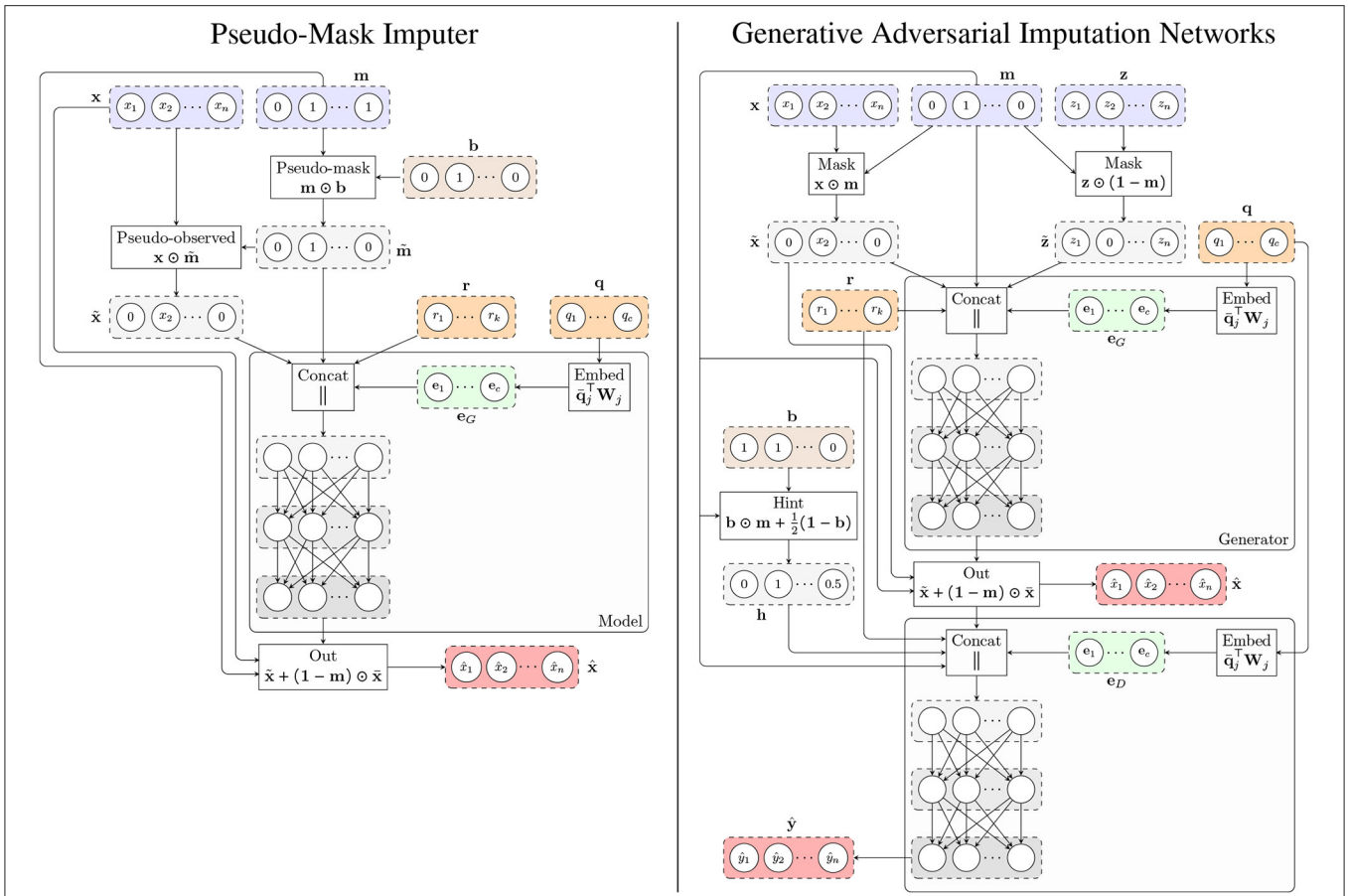

FIGURE 1 | Architecture of the proposed methods. **(Left)** Pseudo-Mask Imputer (PMI). **(Right)** Generative Adversarial Imputation Networks for GTEx (GAIN-GTEx). In both cases, the imputer takes gene expression values $\tilde{\mathbf{x}}$ with missing components according to a mask $\mathbf{m}$, and categorical (e.g., tissue type; $\mathbf{q}$) and numerical (e.g., age; $\mathbf{r}$) covariates, and outputs the imputed values $\bar{\mathbf{x}}$. The observed components of the imputer's output are then replaced by the actual observed expression values $\tilde{\mathbf{x}}$, yielding the imputed sample $\hat{\mathbf{x}}$. For PMI, the *pseudo-mask* mechanism masks out some of the observed components, which are then recovered at the output. For the adversarial model (right), we additionally train a discriminator that receives $\hat{\mathbf{x}}$, the sample covariates, and the hint vector $\mathbf{h}$, and produces the output $\hat{\mathbf{y}}$, whose i-th component $\hat{y}_i$ represents the probability of gene i being observed as opposed to being imputed by the generator.

each $i \in \{1, \ldots, n\}$ and the vector $\mathbf{h} \in \mathbb{R}^n$ corresponds to the *hint* mechanism described in Yoon et al. (2018), which provides theoretical guarantees on the uniqueness of the global minimum for the estimation of $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Concretely, the role of the hint vector $\mathbf{h}$ is to *leak* some information about the mask $\mathbf{m}$ to the discriminator. Similar to GAIN, we define the hint $\mathbf{h}$ as follows:

$$\mathbf{h} = \mathbf{b} \odot \mathbf{m} + \frac{1}{2}(\mathbf{1} - \mathbf{b}) \qquad \mathbf{b} \sim B(1, p) \qquad p \sim U(\alpha, \beta), \quad (11)$$

where $\mathbf{b} \in \{0, 1\}^n$ is a binary vector that controls the amount of information from the mask $\mathbf{m}$ revealed to the discriminator. In contrast to GAIN, which discloses all but one components of the mask, we sample $\mathbf{b}$ from a Bernoulli distribution parametrized by a random probability $p \sim U(\alpha, \beta)$, where $\alpha \in [0, 1]$ and $\beta \in [\alpha, 1]$ are hyperparameters. This accounts for a high number of genes n and allows to trade off the number of mask components that are revealed to the discriminator.

Optimization. Similarly to GAN and GAIN, we optimize the generator and discriminator adversarially, interleaving gradient updates for the discriminator and generator.

The discriminator aims at determining whether genes have been observed or imputed based on the imputed vector $\hat{\mathbf{x}}$, the covariates $\mathbf{q}$ and $\mathbf{r}$, and the hint vector $\mathbf{h}$. Since the hint vector $\mathbf{h}$ readily provides partial information about the ground truth $\mathbf{m}$ (Equation 11), we penalize D only for genes $i \in \{1, 2, \ldots, n\}$ such that $h_i = 0.5$, that is, genes whose corresponding mask value is unavailable to the discriminator. We achieve this via the following loss function $\mathcal{L}_D : \{0, 1\}^n \times \mathbb{R}^n \times \{0, 1\}^n \to \mathbb{R}$:

$$\mathcal{L}_D(\mathbf{m}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z}(\mathbf{1} - \mathbf{b})^\top \big(\mathbf{m} \odot \log \hat{\mathbf{y}} + (\mathbf{1} - \mathbf{m}) \odot (\mathbf{1} - \log \hat{\mathbf{y}})\big), \quad (12)$$

where $Z = 1 + (\mathbf{1} - \mathbf{b})^\top(\mathbf{1} - \mathbf{b})$ is a normalization term. The only difference with respect to the binary cross entropy loss function is the dot product involving $(\mathbf{1} - \mathbf{b})$, which we

employ to ignore genes whose mask has been *leaked* to the discriminator through **h**.

The generator aims at implicitly estimating $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Therefore, its role is not only to impute the expression corresponding to missing genes, but also to reconstruct the expression of the observed inputs. Similar to GAIN, in order to account for this and encourage a realistic imputation function, we use the following loss function $\mathcal{L}_G : \{0,1\}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \{0,1\}^n \rightarrow \mathbb{R}$ for the generator:

$$\mathcal{L}_G(\mathbf{m}, \mathbf{x}, \bar{\mathbf{x}}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z_1}\big((\mathbf{1}-\mathbf{b})\odot(\mathbf{1}-\mathbf{m})\big)^\top \log \hat{\mathbf{y}} + \frac{\lambda}{Z_2}\mathbf{m}^\top(\mathbf{x}-\bar{\mathbf{x}})^2, \tag{13}$$

where $Z_1 = 1+(\mathbf{1}-\mathbf{b})^\top(\mathbf{1}-\mathbf{b})$ and $Z_2 = \mathbf{m}^\top\mathbf{m}$ are normalization terms, and $\lambda > 0$ is a hyperparameter. Intuitively, the first term in Equation (13) corresponds to the adversarial loss, whereas the mean squared error (MSE) term accounts for the loss that the generator incurs in the reconstruction of the observed gene expression values.

Architecture. We model the discriminator D and the generator G using neural networks. Similar to PMI, D and G leverage independent instances $\mathbf{e}^G$ and $\mathbf{e}^D$ of the categorical embeddings described in Equation (6). Specifically, we model the two players as follows:

$$\begin{aligned} G(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \mathbf{m}, \mathbf{r}, \mathbf{q}) &= \text{MLP}(\tilde{\mathbf{x}}\|\tilde{\mathbf{z}}\|\mathbf{m}\|\mathbf{r}\|\mathbf{e}^G) \\ D(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{r}, \mathbf{q}) &= \text{MLP}(\hat{\mathbf{x}}\|\mathbf{h}\|\mathbf{r}\|\mathbf{e}^D), \end{aligned} \tag{14}$$

where MLP denotes a multilayer perceptron and $\tilde{\mathbf{x}} = \mathbf{x}\odot\mathbf{m} \in \mathbb{R}^n$ and $\tilde{\mathbf{z}} = \mathbf{z}\odot(\mathbf{1}-\mathbf{m}) \in \mathbb{R}^n$ are the masked gene expression and noise input vectors, respectively. **Figure 1** shows the architecture of both players.

3. EXPERIMENTAL DETAILS

In this section, we provide an overview of the dataset and describe the experimental details, including all the different case studies and imputation scenarios that we considered. We also describe the implementation details of PMI (see **Supplementary Figure 6**) and GAIN-GTEx (see **Supplementary Figure 7**).

3.1. Materials

Dataset. The GTEx dataset is a public genomic resource of genetic effects on the transcriptome across a broad collection of human tissues, enabling linking of these regulatory mechanisms to trait and disease associations (Aguet et al., 2020). Our dataset contained 15,201 RNA-Seq samples collected from 49 tissues of 838 unique donors. We also selected the intersection of all the protein-coding genes among these tissues, yielding 12,557 unique genes. In addition to the expression data, we leveraged metadata about the sample donors, including sex, age, and cohort (post-mortem, surgical, or organ donor).

Standardization. A large proportion of gene expression data in public repositories contains normalized values. Thus, imputation in this context has practical utility. Imputing the relative expression levels (in normalized data) vs absolute levels (in non-normalized data) is also biologically meaningful, with important applications, e.g., differential expression analysis (between disease individuals and controls) that is robust to expression outliers. To this end, we normalized the expression data via the standard score, so that the standardized expression values have mean 0 and standard deviation 1 for each gene across all samples.

Training, validation, and test splits. To prevent any leakage of information between the training and test sets, we enforced all samples from the same donor to be within the same set. Concretely, we first flipped the GTEx donor identifiers (e.g., 111CU-1826 is flipped to 6281-UC111), we then sorted the reversed identifiers in alphabetical order, and we finally selected a suitable split point, forcing the two sets to be disjoint. After splitting the data, the training set, which we used to train the model, consisted of $\sim$ 60% of the total samples. The validation set, which we used to optimize the method, consisted of $\sim$ 20% of the total samples. The test set, on which we evaluated the final performance, contained the remaining $\sim$ 20% of the data.

3.2. Case Studies

We benchmarked the methods on two case studies:

Case 1: Protein-coding genes. As a first case study, we selected the intersection of all the protein-coding genes among the 49 GTEx tissues, resulting in a set of 12,557 unique genes. This case study is challenging for imputation methods that are not scalable across the number of input variables.

Case 2: Genes in a pathway. We selected a subset of 273 genes from the Alzheimer's disease pathway extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000). This case study allows to benchmark imputation methods that do not scale well with the number of variables.

3.3. Imputation Scenarios

We considered two realistic imputation scenarios:

Scenario 1: In-place imputation. Our goal is to impute the missing values of a dataset $\mathcal{D} = \{(\mathbf{m}\odot\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$ without access to the ground truth missing values $(\mathbf{1}-\mathbf{m})\odot\mathbf{x}$. Importantly, for this scenario we assumed that the data is *missing completely at random* (MCAR; Little and Rubin, 2019), that is, the missingness does not depend on any of the observed nor unobserved variables.

Scenario 2: Inductive imputation. Given a training dataset $\mathcal{D}_{train} = \{(\mathbf{x}, \mathbf{1}, \mathbf{r}, \mathbf{q})\}$ where all expression values $\mathbf{x} \in \mathbb{R}^n$ are observed, our goal is to impute the missing expression values of an independent test dataset $\mathcal{D}_{test} = \{(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$. Methods trained in inductive mode (e.g., on comprehensive datasets such as GTEx) can be used to perform imputation on small, independent datasets where the small number of samples is insufficient to train a model in in-place mode.

3.4. Implementation

For both PMI and GAIN-GTEx, we included the donor's age as numerical covariate in **r** and the tissue type, sex and cohort as categorical covariates in **q**. We normalized the numerical variables via the standard score. For each categorical variable

TABLE 1 | Gene expression imputation performance with a missing rate of 50% across 3 runs (complete set of protein-coding genes).

	Scenario 1: In-place imputation		Scenario 2: Inductive imputation	
Method	R^2	**Runtime (hours)**	R^2	**Runtime (hours)**
MICE	–	–	–	–
MissForest	–	–	–	–
Blood surrogate	-0.693 ± 0.000	0.000 ± 0.000	-0.952 ± 0.000	0.000 ± 0.000
Median imputation	0.000 ± 0.000	0.001 ± 0.000	-0.009 ± 0.000	0.001 ± 0.000
1-NN imputation	0.179 ± 0.000	1.616 ± 0.004	0.203 ± 0.000	0.985 ± 0.003
5-NN imputation	0.461 ± 0.000	2.224 ± 0.107	0.482 ± 0.000	1.441 ± 0.096
10-NN imputation	0.468 ± 0.000	2.140 ± 0.035	0.495 ± 0.000	1.711 ± 0.160
GAIN-MSE-GTEx	0.637 ± 0.005	0.199 ± 0.074	0.638 ± 0.003	0.456 ± 0.053
GAIN-GTEx	0.638 ± 0.007	0.625 ± 0.294	0.636 ± 0.001	1.199 ± 0.157
PMI	0.479 ± 0.003	0.241 ± 0.024	0.707 ± 0.001	0.244 ± 0.019

We do not report the R^2 scores for MICE and MissForest, because the runtime is longer than 7 days. Note GAIN-GTEx outperforms all the other methods in in-place imputation while PMI displays the highest performance in inductive imputation.

$q_j \in \{1, 2, \ldots, v_j\}$, we used the rule of thumb $d_j = \lfloor \sqrt{v_j} \rfloor + 1$ to set all the dimensions of the categorical embeddings. We used ReLU activations for each hidden layer in the MLP architectures of both PMI and GAIN (see Equations 7 and 14).

We trained both models using the Adam optimizer (Kingma and Ba, 2014). We used batch normalization (Ioffe and Szegedy, 2015) in the hidden layers of the MLPs, which yielded a significant speed-up to the training convergence according to our experiments. We used early stopping with a patience of 30. The rest of parameters for each model, case study, and imputation scenario are presented in the **Supplementary Material**.

3.5. Baseline Methods

We compared PMI and GAIN-GTEx to several baseline methods:

Common methods of imputation. We considered two simple gene expression imputation approaches: blood surrogate and median imputation. The use of blood, an easily accessible tissue, as a surrogate for difficult-to-acquire tissues is done in studies of biomarker discovery, diagnostics, and eQTLs, and in the development of model systems (Gamazon et al., 2018; Kim et al., 2020). For blood surrogate imputation, we imputed missing gene expression values in any given tissue with the corresponding values in whole blood for the same donor. For median imputation, we imputed missing values with the median of the observed tissue-specific gene expression computed across donors.

***k*-Nearest Neighbours**. The *k*-Nearest Neighbours (*k*-NN) algorithm is an efficient method that is commonly used for imputation (Beretta and Santaniello, 2016). Here, we leveraged *k*-NN as a baseline for different values of *k*. This model estimates the missing values of a sample based on the values of the missing components in the *k* closest samples.

State-of-the-art methods. We considered two state-of-the-art imputation methods: Multivariate Imputation by Chained Equations (MICE; Buuren and Groothuis-Oudshoorn, 2010) and MissForest (Stekhoven and Bühlmann, 2012). MICE leverages chained equations to create multiple imputations of missing data. The hyperparameters of MICE include the minimum/maximum possible imputed value for each component and the maximum number of imputation rounds. MissForest (Stekhoven and Bühlmann, 2012) is a non-parametric imputation method based on random forests trained on observed values to impute the missing values. Among others, the hyperparameters of MissForest include the number of trees in the forest and the number of features to consider when looking for the optimal split.

4. RESULTS

Here we provide an overview of the imputation results, including a comparison with other imputation methods, an evaluation of the tissue-specific results, and an analysis of the cross-study relevance across different levels of missingness.

4.1. Comparison

Tables 1, **2** show a quantitative summary of the imputation performances for the two case-studies and the two imputation scenarios. In addition to the imputation scores, we report the runtime of all the methods. We labeled methods as computationally *unfeasible* when they took longer than 7 days to run on our server (CPU: Intel(R) Xeon(R) Processor E5-2630 v4. RAM: 125GB), after which we halted the execution. For example, MICE and MissForest were unfeasible for each imputation scenario on the complete set of protein-coding genes. An empirical study of the scalability of both methods (see **Supplementary Material**) showed that the runtime increases rapidly with the number of genes. However, on a smaller set of genes (i.e., 273 from the Alzheimer's disease pathway), evaluation of the performance was successfully obtained, with the runtime substantially higher for both methods than for the other methods. In addition, we included GAIN-MSE-GTEx as a baseline, consisting of a simplification of GAIN-GTEx that was optimized exclusively via the mean squared error term of the generator. GAIN-MSE-GTEx performed reasonably well relative to GAIN-GTEx, suggesting that the mean squared error term of the loss function was driving the learning (see **Supplementary Material**).

TABLE 2 | Gene expression imputation performance with a missing rate of 50% across 3 runs (for a subset of 273 genes from the Alzheimer's disease pathway).

	Scenario 1: In-place imputation		Scenario 2: Inductive imputation	
Method	R^2	**Runtime (hours)**	R^2	**Runtime (hours)**
MICE	0.574 ± 0.001	2.062 ± 0.335	0.569 ± 0.001	2.252 ± 0.096
MissForest (1 tree)	-0.147 ± 0.002	0.145 ± 0.002	-0.042 ± 0.003	0.575 ± 0.167
MissForest (10 trees)	0.458 ± 0.001	0.839 ± 0.176	0.514 ± 0.001	3.220 ± 0.371
MissForest (20 trees)	0.478 ± 0.000	1.836 ± 0.068	0.540 ± 0.000	4.842 ± 0.495
MissForest (100 trees)	0.493 ± 0.000	6.438 ± 0.498	0.561 ± 0.001	16.186 ± 1.709
Blood surrogate	-0.698 ± 0.002	0.000 ± 0.000	-0.971 ± 0.002	0.000 ± 0.000
Median imputation	0.001 ± 0.000	0.000 ± 0.000	-0.009 ± 0.000	0.000 ± 0.000
1-NN imputation	0.186 ± 0.001	0.037 ± 0.001	0.301 ± 0.000	0.021 ± 0.001
GAIN-MSE-GTEx	0.519 ± 0.001	0.038 ± 0.002	0.533 ± 0.001	0.045 ± 0.004
GAIN-GTEx	0.533 ± 0.001	0.139 ± 0.041	0.527 ± 0.003	0.569 ± 0.017
PMI	0.536 ± 0.001	0.048 ± 0.002	0.630 ± 0.011	0.037 ± 0.002

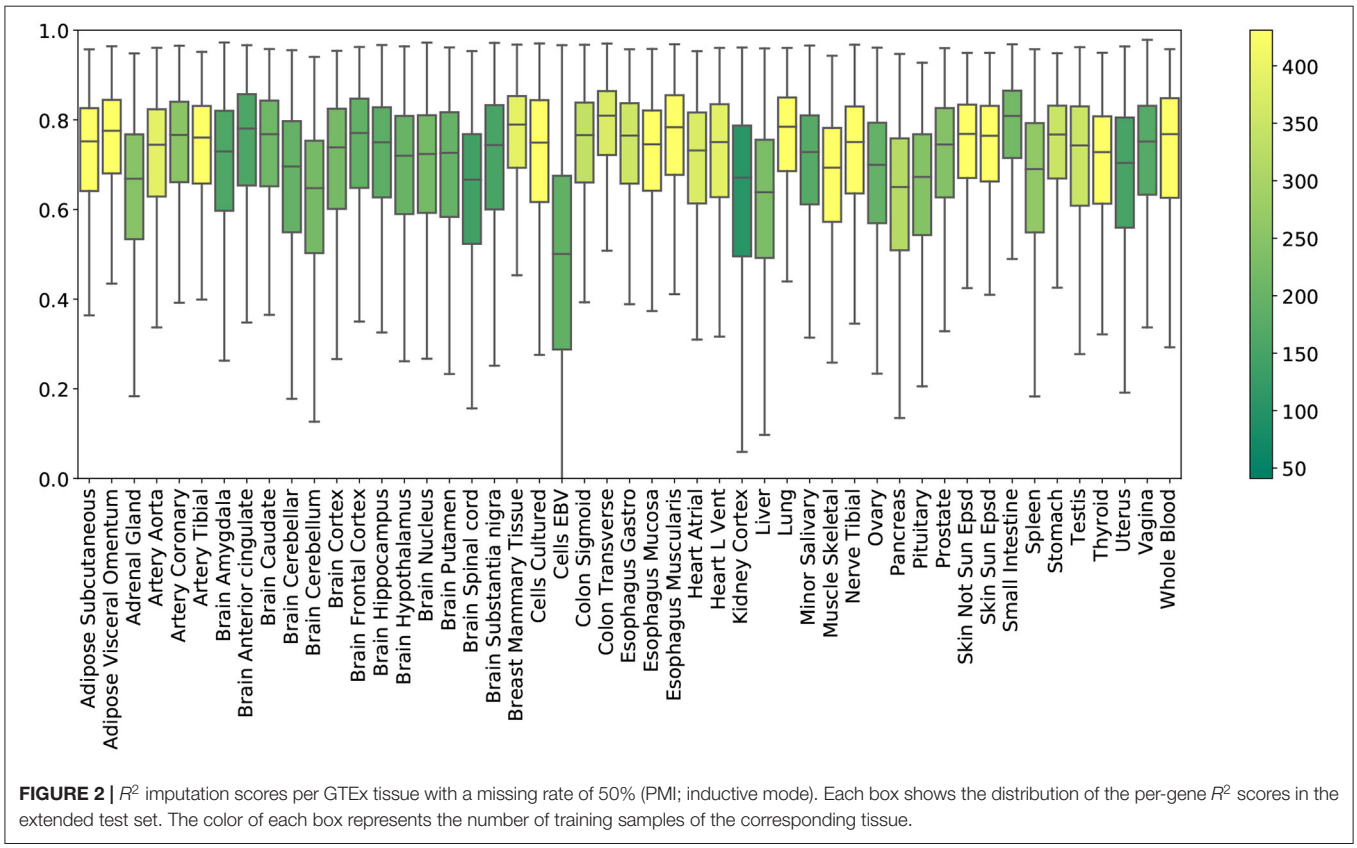

FIGURE 2 | R^2 imputation scores per GTEx tissue with a missing rate of 50% (PMI; inductive mode). Each box shows the distribution of the per-gene R^2 scores in the extended test set. The color of each box represents the number of training samples of the corresponding tissue.

In terms of the evaluation metrics, we report the coefficient of determination (R^2). This metric ranges from $-\infty$ to 1 and corresponds to the ratio of explained variance to the total variance. Negative scores indicate that the model predictions are worse than those of a baseline model that predicts the mean of the data. Here, to evaluate the performance, we generated random masks with a missing rate of 50% and computed the imputation R^2 per gene. We repeated the last step 3 times and reported the overall mean R^2 and the average per-gene standard deviation of the R^2 scores, averaged across the 3 runs. In inductive mode, blood surrogate and median imputation exhibited negative scores. Under in-place imputation on the protein-coding genes, GAIN-GTEx outperformed all the other methods (0.638 ± 0.007). Under inductive imputation on the protein-coding genes, PMI showed the best overall performance (0.707 ± 0.001) among all the methods.

4.2. Imputation Results

Tissue-specific results. **Figure 2** shows the R^2 scores achieved by PMI across all 49 tissue types. To obtain these results, we

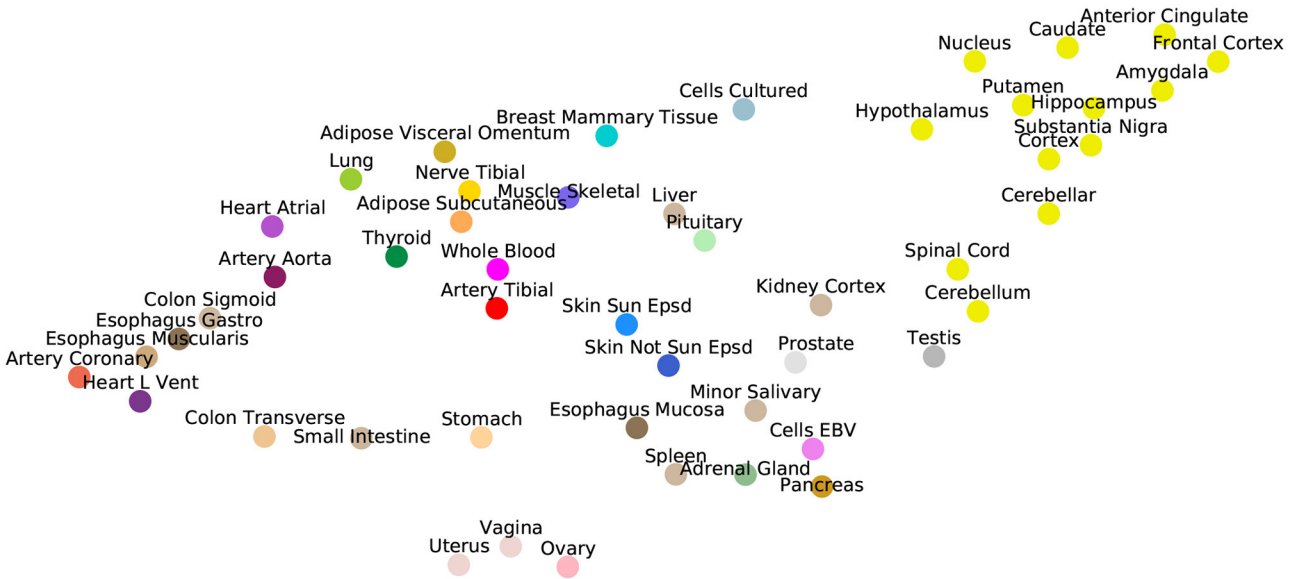

FIGURE 3 | UMAP visualization of the tissue embeddings from the generator. Colors are assigned to conform to the GTEx Consortium conventions. Note that the central nervous system, consisting of the 13 brain regions, clusters together on the top right corner.

TABLE 3 | Cross-study results for GAIN-GTEx and PMI trained on GTEx (inductive mode).

GAIN-GTEx		PMI	
Tissue	R^2	**Tissue**	R^2
TCGA LAML	0.386 ± 0.057	TCGA LAML	0.394 ± 0.065
TCGA BRCA	0.408 ± 0.023	TCGA BRCA	0.427 ± 0.023
TCGA LUAD	0.439 ± 0.034	TCGA LUAD	0.451 ± 0.050
GTEx Whole blood	0.678 ± 0.031	GTEx Whole blood	0.709 ± 0.034
GTEx Breast	0.724 ± 0.036	GTEx Breast	0.751 ± 0.039
GTEx Lung	0.713 ± 0.033	GTEx Lung	0.744 ± 0.035

We report the R^2 scores on data from 3 TCGA cancer types and their healthy counterpart on GTEx for a missing rate of 50%.

generated random masks with a missing rate of 50% for the test set, performed imputation, and plotted the distribution of 12,557 gene R^2 scores for each tissue. Mean R^2 scores in the individual tissues ranged from $\sim$ 0.5 (Epstein Barr virus transformed lymphocytes; EBV) to $\sim$ 0.78 (small intestine). Kidney cortex, the tissue with the smallest sample size, had the highest variability in R^2 with an interquartile range of $Q_3 - Q_1 = 0.30$.

Figure 3 illustrates the ability of GAIN-GTEx to learn rich tissue representations. Specifically, we plotted a UMAP representation (McInnes et al., 2018) of the learnt tissue embeddings $\mathbf{W}_j \in \mathbb{R}^{49\times 8}$ from the generator (see Equation 5), where j indexes the tissue dimension. Strikingly, the tissue representations showed strong clustering of biologically-related tissues, including the central nervous system (i.e., the 13 brain regions), the gastrointestinal system (e.g., the esophageal and colonic tissues), and the female reproductive tissues (i.e., uterus, vagina, and ovary). The clustering properties were robust across UMAP runs and could be similarly appreciated using other dimensionality reduction algorithms such as tSNE (Maaten and Hinton, 2008).

Cross-study results across missing rates. To evaluate the cross-study relevance and generalizability of PMI and GAIN-GTEx, we leveraged the model trained on GTEx to perform imputation on The Cancer Genome Atlas (TCGA) gene expression data in acute myeloid leukemia (TCGA LAML; Cancer Genome Atlas Research Network et al., 2013), breast cancer (TCGA BRCA; Cancer Genome Atlas Network, 2012), and lung adenocarcinoma (TCGA LUAD; Cancer Genome Atlas Research Network, 2014). For each TCGA tissue and its *non-diseased* test counterpart on GTEx, we show the imputation quality in **Table 3** as well as the performance across varying missing rates in **Figure 4**.

Imputation results on genes from the Alzheimer's disease pathway. **Figure 5** shows the per-gene imputation scores for GAIN-GTEx trained on a subset of 273 genes corresponding to the Alzheimer's disease pathway. Amyloid-beta is a core element of senile plaques which are characteristic of the debilitating disease, with various pathophysiological consequences on cellular processes. The pathway consists of genes that are involved in a number of processes, including neuronal apoptosis, autophagy deficits, mitochondrial defect, and neurodegeneration. We observed that some genes in the pathway (e.g., PSMB6, COX6C, PSMD7, PSMA2, PSMD14, SDHB, TUBB1, TUBA8, FZD9, LPL, KIF5C, TUBB4A, TUBB2B, APOE) exhibited different distributions between brain and non-brain tissue types. The most highly imputed genes were enriched in known gene sets (see **Supplementary Figures 9, 10**).

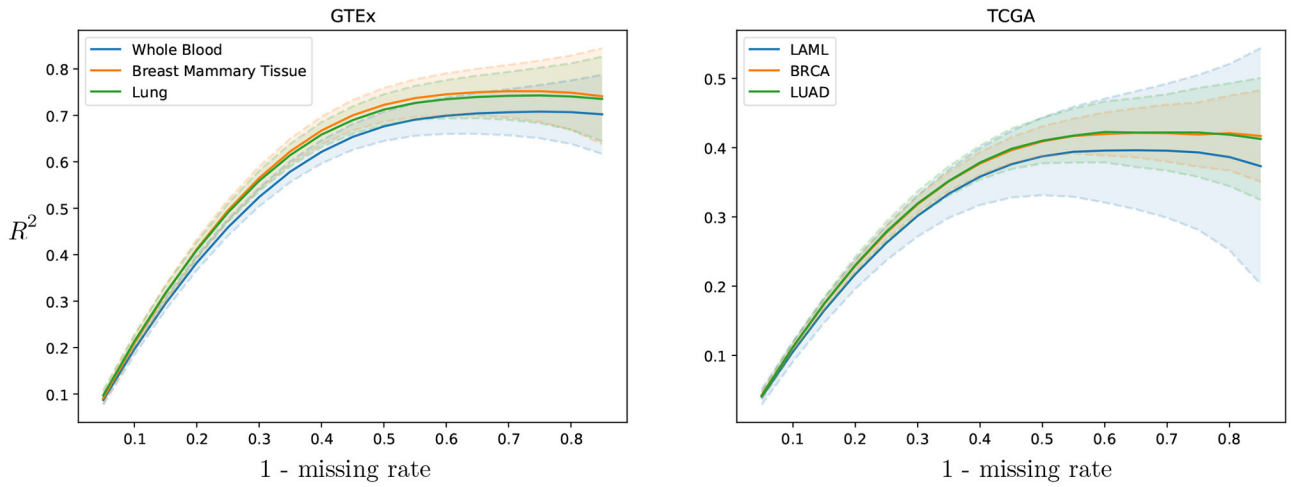

FIGURE 4 | GAIN-GTEx R^2 imputation scores per tissue across missing rate for 3 TCGA cancer types and their healthy counterpart in GTEx. The shaded area represents one standard deviation of the per-gene R^2 scores in the corresponding tissue. The greater the rate of missingness, the lower the performance.

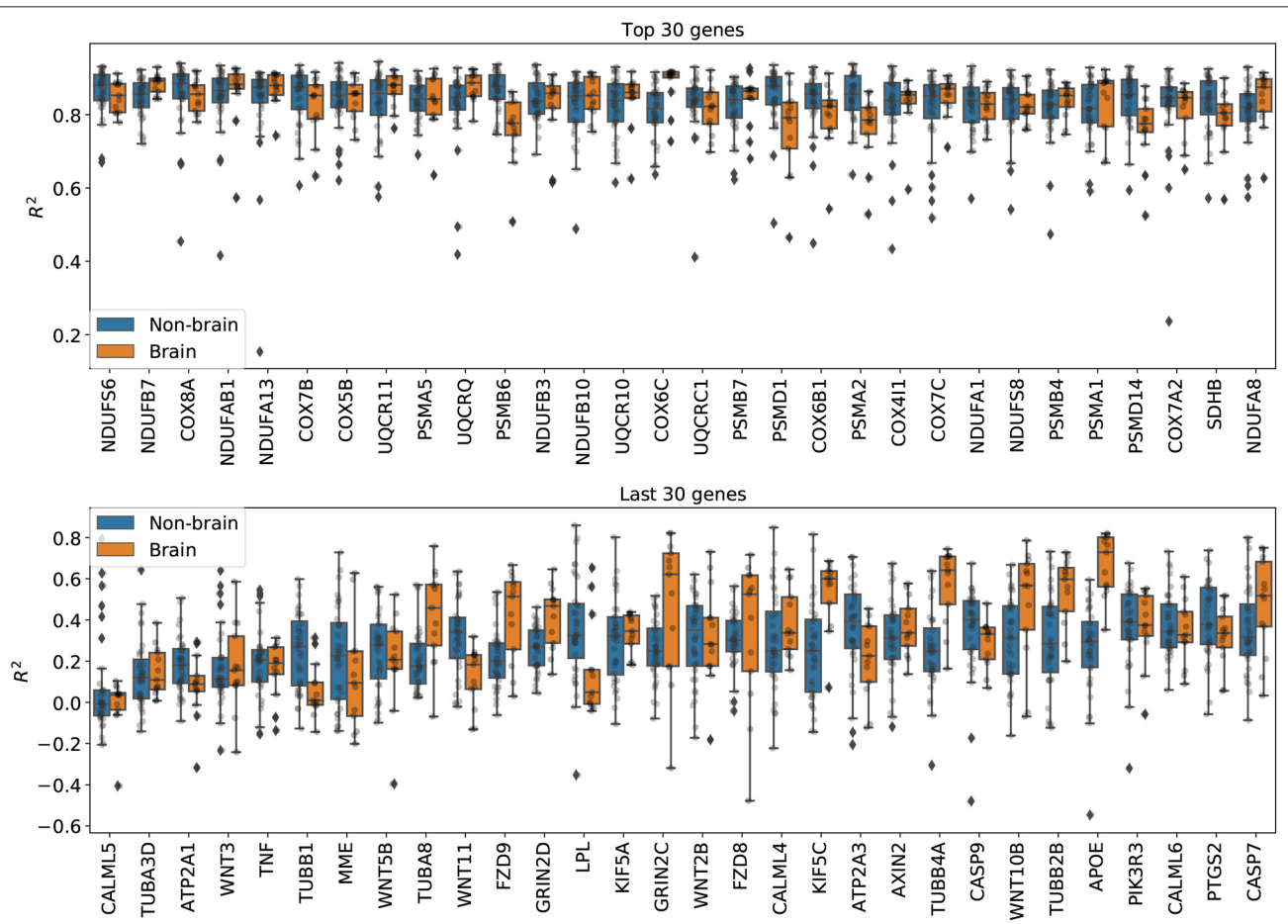

FIGURE 5 | Per-gene imputation R^2 scores on genes from the Alzheimer's disease pathway. Each point represents the average R^2 score in a tissue type. We note that some genes in the pathway (e.g., PSMB6, COX6C, PSMD7, PSMA2, PSMD14, SDHB, TUBB1, TUBA8, FZD9, LPL, KIF5C, TUBB4A, TUBB2B, APOE) exhibited different distributions between brain and non-brain tissue types.

5. DISCUSSION

We developed two imputation approaches to gene expression, facilitating the reconstruction of a high-dimensional molecular trait that is central to disease biology and drug target discovery. The proposed methods, which we called Pseudo-Mask Imputer (PMI) and GAIN-GTEx, were able to approximate the gene expression manifold from incomplete gene expression data and relevant covariates (potential global determinants of expression) and impute missing expression values. A characteristic feature of our architectures is the use of word embeddings, which enabled to learn distributed representations of the tissue types (see **Figure 3**). Importantly, this allowed to condition the imputation algorithms on factors that drive gene expression, endowing the architectures with the ability to represent them in a biologically meaningful way.

We leveraged the most comprehensive human transcriptome resource available (GTEx), allowing us to test the performance of our method in a broad collection of tissues (see **Figure 2**). The biospecimen repository includes commonly used surrogate tissues (whole blood and EBV transformed lymphocytes), central nervous system tissues from 13 brain regions, and a wide diversity of other primary tissues from *non-diseased* individuals. In particular, we observed that EBV transformed lymphocytes, an accessible and renewable resource for functional genomics, are a notable outlier in imputation performance. This is perhaps not surprising, consistent with studies about the transcriptional effect of EBV infection on the suitability of the cell lines as a model system for primary tissues (Carter et al., 2002). Interestingly, similar tissues exhibit similar R^2 scores (see **Supplementary Figure 12**).

We analyzed the performance of the proposed approaches and found that they compare favorably to several existing imputation methods in terms of imputation performance and runtime (see **Table 1**). We observed that standard approaches such as leveraging the expression of missing genes from a surrogate blood tissue yielded negative R^2 values and therefore did not perform well. Median imputation, although easy to implement, had a very limited predictive power. Imputation methods based on k-Nearest Neighbours were computationally feasible and yielded solid but poorer R^2 scores. In terms of state-of-the-art-methods, MICE and MissForest were computationally prohibitive given the high-dimensionality of the data and we halted the execution after running our experiments for 7 days. In particular, we performed an empirical study of the scalability of both methods (see **Supplementary Figures 1–5**) and observed that the runtime increases very rapidly with the number of genes. To alleviate this issue, we compared PMI and GAIN-GTEx with these methods on a subset of 273 genes from the Alzheimer's disease pathway (see **Table 2**). Under the in-place imputation scenario (Alzheimer's disease pathway), MICE performed better than PMI, GAIN-GTEx, and MissForest (100 trees). Under the inductive imputation setting, PMI outperformed all the other methods by a large margin.

In terms of the comparison between PMI and GAIN-GTEx, our experiments suggest that the latter is generally harder to optimize (see hyperparameter search in **Supplementary Material**). In particular, GAIN resembles a deep autoencoder in that the supervised loss penalizes the reconstruction error of the observed components. While this is a natural choice, autoencoder-like architectures are considerably sensitive to the user-definable bottleneck dimension. On one hand, a small number of units results in under-fitting. On the other hand, an excessively big bottleneck dimension allows the neural network to trivially *copy-paste* the observed components. In contrast, the loss function of PMI does not penalize the reconstruction error for the *pseudo-observed* components (e.g., the loss function of PMI penalizes the prediction error of the *pseudo-missing* components, which are not provided as input at training time). Together with the fact that the *pseudo-mask* mechanism dynamically enlarges the training size, this subtlety allows training considerably bigger networks without over-fitting. Finally, we observed that a simplification of GAIN-GTEx, GAIN-MSE-GTEx, performed similarly well, suggesting that the mean squared error term of the generator's loss function is driving the learning process. In **Supplementary Material**, we discuss our empirical findings about the adversarial loss of GAIN. For the purpose of reproducibility, as the gains of the adversarial loss appear to be small or negligible given our observations, we recommend training GAIN-GTEx without the adversarial term.

To evaluate the cross-study relevance of our method, we applied the trained models derived from GTEx (inductive mode) to perform imputation on The Cancer Genome Atlas gene expression data in acute myeloid leukemia, lung adenocarcinoma, and breast cancer. In addition to technical artifacts (e.g., batch effects), generalizing to this data is highly challenging because the expression is largely driven by features of the disease such as chromosomal abnormalities, genomic instabilities, large-scale mutations, and epigenetic changes (Baylin and Jones, 2011; Weinstein et al., 2013). Our results show that, despite these challenges, the methods were robust to gene expression from multiple diseases in different tissues (see **Table 3**), lending themselves to being used as tools to extend independent transcriptomic studies. Next, we evaluated the imputation performance of PMI and GAIN-GTEx for a range of values for the missing rate (see **Figure 4** and **Supplementary Figure 8**). We noted that the performance is stable and that the greater the proportion of missing values, the lower the prediction performance. Finally, we analyzed the imputation performance across genes from the Alzheimer's disease pathway (see **Figure 5**) and across all genes (see **Supplementary Figure 9**). We observed that the most highly imputed imputed genes are non-random and, indeed, cluster in some known pathways (see **Supplementary Figures 10, 11**).

Broader Impact. The study of the transcriptome is fundamental to our understanding of cellular and pathophysiological processes. High-dimensional gene expression data contain information relevant to a wide range of applications, including disease diagnosis (Huang et al., 2010), drug development (Sun et al., 2013), and evolutionary inference (Colbran et al., 2019). Thus, accurate and robust methods for imputation of gene expression have the enormous potential to enhance our molecular understanding of complex diseases, inform the search for novel drugs, and provide key insights into evolutionary processes. Here, we developed a

methodology that attains state-of-the art performance in several scenarios in terms of imputation quality and execution time. Our analysis showed that the use of blood as a surrogate for difficult-to-acquire tissues, as commonly practized in biomedical research, may lead to substantially degraded performance, with important implications for biomarker discovery and therapeutic development. Our method generalizes to gene expression in a disease class which has shown considerable health outcome disparities across population groups in terms of morbidity and mortality. Future algorithmic developments therefore hold promise for more effective detection, diagnosis, and treatment (Hosny and Aerts, 2019) and for improved implementation in clinical medicine (Char et al., 2018). Increased availability of transcriptomes in diverse human populations to enlarge our training data (a well-known and critical ethical challenge) should lead to further gains (i.e., decreased biases in results and reduced health disparities) (Wojcik et al., 2019). This work has the potential to catalyze research into the application of deep learning to molecular reconstruction of cellular states and downstream gene mapping of complex traits (Cookson et al., 2009; Zhou et al., 2020).

6. CONCLUSION

In this work, we developed two methods for gene expression imputation, which we named PMI and GAIN-GTEx. To increase the applicability of the proposed methods, we trained them on RNA-Seq data from the Genotype-Tissue Expression project, a reference resource that has generated a comprehensive collection of transcriptomes in a diverse set of tissues. A characteristic feature of our architectures is the use of word embeddings to learn distributed representations for the tissue types. Our approaches compared favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime, and generalized to transcriptomics data from 3 cancer types of the The Cancer Genome Atlas. PMI and GAIN-GTEx show optimal performance among the methods in inductive and in-place imputation, respectively, on the protein-coding genes. This work can facilitate the straightforward integration and cost-effective repurposing of large-scale RNA biorepositories into genomic studies of disease, with high applicability across diverse tissue types.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the GTEx portal: https://gtexportal.org/. A detailed summary of the GTEx samples and donor information can be found at: https://gtexportal.org/home/tissueSummaryPage. Our code is publicly available at https://github.com/rvinas/GTEx-imputation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RV and TA developed and trained the deep learning algorithm, generated all the results and figures. ERG provided the standardized RNA-seq data. ERG and PL supervised the study as joint senior authors. All the authors wrote and approved the manuscript.

FUNDING

The project leading to these results has received funding from la Caixa Foundation (ID 100010434), under agreement LCF/BQ/EU19/11710059. This research was supported by the National Institutes of Health under award numbers R35HG010718 (ERG), R01HG011138 (ERG), R01GM140287 (ERG), and R01HL133559 (ERG). This research was also funded by the W. D. Armstrong Trust Fund, University of Cambridge, UK (TA) and the Engineering and Physical Sciences Research Council (R.V. EPSRC DTG 2018/19). PL was supported by MICA: Mental Health Data Pathfinder: University of Cambridge, Cambridgeshire and Peterborough NHS Foundation Trust, Microsoft, and the Medical Research Council (MC_PC_17213).

ACKNOWLEDGMENTS

We thank Nikola Simidjievski, Cătălina Cangea, Ben Day, Cristian Bodnar, and Arian Jamasb for the helpful comments and discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624128/full#supplementary-material

REFERENCES

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., et al. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1101/787903

Baylin, S. B., and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer* 11, 726–734. doi: 10.1038/nrc3130

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med. Inform. Decis. Mak.* 16:74. doi: 10.1186/s12911-016-0318-z

Buuren, S. V., and Groothuis-Oudshoorn, K. (2010). mice: multivariate imputation by chained equations in r. *J. Stat. Softw.* 45, 1–68. doi: 10.18637/jss.v045.i03

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490:61. doi: 10.1038/nature11412

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/nature13385

Cancer Genome Atlas Research Network, Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., et al. (2013). Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* 368, 2059–2074. doi: 10.1056/NEJMoa1301689

Carter, K. L., Cahir-McFarland, E., and Kieff, E. (2002). Epstein-barr virus-induced changes in b-lymphocyte gene expression. *J. Virol.* 76, 10427–10436. doi: 10.1128/JVI.76.20.10427-10436.2002

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* 378:981. doi: 10.1056/NEJMp1714229

Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. (2019). Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* 3, 1598–1606. doi: 10.1038/s41559-019-0996-x

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-1047-4

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194. doi: 10.1038/nrg2537

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423–428. doi: 10.1038/nature06758

Evans, W. E., and Relling, M. V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, 464–468. doi: 10.1038/nature02626

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nat. Genet.* 50, 956–967. doi: 10.1038/s41588-018-0154-4

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems* - Volume 2 (Cambridge, MA: MIT Press), 2672-2680.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Gupta, A., and Zou, J. (2019). Feedback gan for DNA optimizes protein functions. *Nat. Mach. Intell.* 1, 105–111. doi: 10.1038/s42256-019-0017-4

Hosny, A., and Aerts, H. J. (2019). Artificial intelligence for global health. *Science* 366, 955–956. doi: 10.1126/science.aay5189

Huang, H., Liu, C.-C., and Zhou, X. J. (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6823–6828. doi: 10.1073/pnas.0912043107

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 448–456. Available online at: JMLR.org.

Kanehisa, M., and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kim, K., Kim, M.-J., Kim, D. W., Kim, S. Y., Park, S., and Park, C. B. (2020). Clinically accurate diagnosis of alzheimer's disease via multiplexed sensing of core biomarkers in human plasma. *Nat. Commun.* 11, 1–9. doi: 10.1038/s41467-019-13901-z

King, M.-C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]* arXiv:1412.6980.

Little, R. J., and Rubin, D. B. (2019). *Statistical Analysis With Missing Data*, Vol. 793. New York, NY: John Wiley & Sons.

Low, L. A., Mummery, C., Berridge, B. R., Austin, C. P., and Tagle, D. A. (2020). Organs-on-chips: into the next decade. *Nat. Rev. Drug Discov.*, 1–17. doi: 10.1038/s41573-020-0079-3

Maaten, L. V. D. and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: https://www.jmlr.org/papers/v9/vandermaaten08a.html

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., and Bonn, S. (2020). Realistic *in silico* generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-019-14018-z

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: uniform manifold approximation and projection. *J. Open Sour. Softw.* 3:861. doi: 10.21105/joss.00861

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, Vol. 26, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.). Available online at: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Øystein Sørensen, Hellton, K. H., Frigessi, A., and Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. *J. Comput. Graph. Stat.* 27, 739–749. doi: 10.1080/10618600.2018.1425626

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77. doi: 10.1126/scitranslmed.3001318

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7:500. doi: 10.1038/nprot.2011.457

Stekhoven, D. J., and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Sun, X., Vilar, S., and Tatonetti, N. P. (2013). High-throughput methods for combinatorial drug discovery. *Sci. Transl. Med.* 5:205rv1. doi: 10.1126/scitranslmed.3006667

The GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* 17, 520–525. doi: 10.1093/bioinformatics/17.6.520

Viñas, R., Andrés-Terré, H., Liò, P., and Bryson, K. (2021). Adversarial generation of gene expression data. *Bioinformatics* btab035. doi: 10.1093/bioinformatics/btab035

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*, eds D. P. Berrar, W. Dubitzky, and M. Granzow (Boston, MA: Springer), 91–109.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45:1113. doi: 10.1038/ng.2764

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi: 10.1038/s41586-019-1310-4

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). GAIN: missing data imputation using generative adversarial nets. *arXiv [Preprint]* arXiv:1806.02920.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4. doi: 10.2202/1544-6115.1128

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and mendelian randomization analysis. *Nat. Genet.* 52, 1239–1246. doi: 10.1038/s41588-020-0706-2

Conflict of Interest: ERG receives an honorarium from the journal *Circulation Research* of the American Heart Association, as a member of the Editorial Board.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.