A Appendix

A.1 Acknowledgements

SW was supported by the Royal Society of Edinburgh (RSE) (grant number 69938). BRH and JW acknowledge the receipt of studentship awards from the Health Data Research UK & The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z).

A.2 Data and Code Availability

All data used in this manuscript are publicly available. Gene expression data is version 2 of the adjusted pan-cancer gene expression data obtained from Synapse and can be found at https://www.synapse.org/#!Synapse:syn4976369.2. The first transcriptomic dataset is comprised of the LGG and GBM cancer subtypes in this pan-cancer gene expression dataset and the second is comprised of the LUAD and LUSC cancer subtypes from the same pan-cancer set.

Code is available at https://github.com/roskamsh/BetaVAEMultImpute.

A.3 Software Requirements

The analyses carried out in this manuscript require the following software: python v3.10, TensorFlow v2.7.0; R: penalized v0.9, MASS v7.3, caret v6.0.

A.4 β-VAEs and the Power Likelihood

In the following, we show that the variational parameters $\boldsymbol{\phi}$ which maximize the β-VAE bound equivalently minimize the KL divergence between the variational posterior $q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})$ and the true posterior under the power likelihood. Specifically, the KL divergence between the variational posterior and the posterior under the power likelihood is given by:

$$
\begin{aligned}
D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}), p_{\boldsymbol{\theta},\beta}(\mathbf{Z}|\mathbf{X})) &= E_{\mathbf{Z}\sim q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log\left(\frac{p_{\boldsymbol{\theta},\beta}(\mathbf{X})q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}{p_{\boldsymbol{\theta}}(\mathbf{X}\mid\mathbf{Z})^{1/\beta}p(\mathbf{Z})}\right)\right] \\
&= -\sum_{n=1}^{N} E_{\mathbf{z}_n\sim q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)}\left[\log\left(p_{\boldsymbol{\theta}}(\mathbf{x}_n\mid\mathbf{z}_n)^{1/\beta}\right)\right] \\
&\quad + \sum_{n=1}^{N} E_{\mathbf{z}_n\sim q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)}\left[\log\left(\frac{q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)}{p(\mathbf{z}_n)}\right)\right] + \log\left(p_{\boldsymbol{\theta},\beta}(\mathbf{X})\right) \\
&= \text{const.} - \sum_{n=1}^{N} E_{\mathbf{z}_n\sim q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)}\left[\log\left(p_{\boldsymbol{\theta}}(\mathbf{x}_n\mid\mathbf{z}_n)^{1/\beta}\right)\right] \\
&\quad + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n), p(\mathbf{z}_n)).
\end{aligned}
$$

Thus, we can equivalently find $\boldsymbol{\phi}$, which maximize the ELBO:

$$
\mathrm{ELBO} = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n\sim q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_n|\mathbf{z}_n)] - \beta\, D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n), p(\mathbf{z}_n))
$$

Example: factorized Gaussian. Assume the generative model is a factorized Gaussian (as is used for the genomic data in Section **??**):

$$p_{\boldsymbol{\theta}}(\mathbf{x}_n|\mathbf{z}_n) = \prod_{d=1}^{D} \mathrm{N}\left(x_{n,d} \mid \mu_d(\mathbf{z}_n), \sigma_d^2(\mathbf{z}_n)\right),$$

where $(\mu_d(\mathbf{z}_n), \sigma_d^2(\mathbf{z}_n))$ for $d = 1, \ldots, D$ represent the output of the final layer of the neural network with weights and biases contained in $\boldsymbol{\theta}$. In this case, the full conditional of the missing data under the power likelihood is

$$\begin{aligned}
p_{\boldsymbol{\theta},\beta}(\mathbf{x}_{\mathrm{mis},n} \,|\mathbf{x}_{\mathrm{obs},n}, \mathbf{z}_n) &\propto p_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{mis},n} \mid \mathbf{x}_{\mathrm{obs},n}, \mathbf{z}_n)^{1/\beta} \\
&= \left(\prod_{d \in \mathcal{D}_{\mathrm{mis},n}} \mathrm{N}\left(x_{n,d} \mid \mu_d(\mathbf{z}_n), \sigma_d^2(\mathbf{z}_n)\right) \right)^{1/\beta} \\
&= \left(\prod_{d \in \mathcal{D}_{\mathrm{mis},n}} \frac{1}{\sqrt{2\pi\sigma_d^2(\mathbf{z}_n)}} \exp\left(\frac{1}{2\sigma_d^2(\mathbf{z}_n)} (x_{n,d} - \mu_d(\mathbf{z}_n))^2 \right) \right)^{1/\beta} \\
&\propto \prod_{d \in \mathcal{D}_{\mathrm{mis},n}} \exp\left(\frac{1}{2\beta\sigma_d^2(\mathbf{z}_n)} (x_{n,d} - \mu_d(\mathbf{z}_n))^2 \right) \\
&\propto \prod_{d \in \mathcal{D}_{\mathrm{mis},n}} \mathrm{N}\left(x_{n,d} \mid \mu_d(\mathbf{z}_n), \beta\sigma_d^2(\mathbf{z}_n)\right),
\end{aligned}$$

where $\mathcal{D}_{\mathrm{mis},n} \subseteq \{1, \ldots, D\}$ contains the indices of the missing features for the nth data point. Thus, in this case, sampling from the full conditional of the missing data under the power likelihood corresponds to sampling from the Gaussian with variance rescaled by a factor of β. Note that for $\beta > 1$ this corresponds to increasing the spread and uncertainty of the missing data, which is critical to improve coverage of the deep generative model.

A.5 Sample Importance Resampling

We first note that our SIR scheme differs slightly from the scheme proposed by [1], who propose joint importance samples $(\mathbf{z}_n^{(s)}, \mathbf{x}_{\mathrm{mis},n}^{(s)})$ from

$$\mathbf{z}_n^{(s)} \sim q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_{\mathrm{mis},n}^{(0)}, \mathbf{x}_{\mathrm{obs},n}), \quad \mathbf{x}_{\mathrm{mis},n}^{(s)} \sim p_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{mis},n} \mid \mathbf{x}_{\mathrm{obs},n}, \mathbf{z}_n^{(s)}).$$

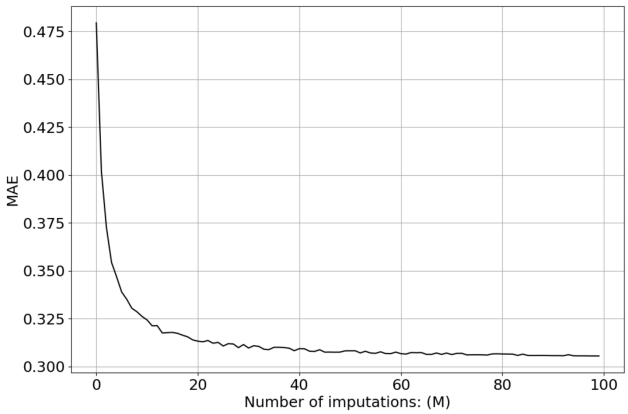
Instead, we only consider importance sampling for $(\mathbf{z}_n^{(s)})$ and subsequently sample the missing data for each of resampled latent variables. Importantly, if the effective sample size is low, resulting in potential duplicates in the M samples of latent variables, we obtain improved variability across multiple imputations of the missing data, compared to the approach of [1].

Example: factorized Gaussian. Assume the generative model is a factorized Gaussian, then the importance weights are proportional to :

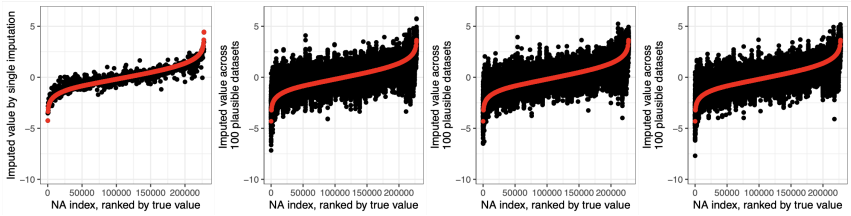
$$\begin{aligned}\omega_n^{(s)} &= \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{obs},n}|\mathbf{z}_n^{(s)})^{1/\beta} p(\mathbf{z}_n^{(s)})}{q_{\boldsymbol{\phi}}(\mathbf{z}_n^{(s)}|\mathbf{x}_{\text{mis},n}^{(0)}, \mathbf{x}_{\text{obs},n})} \\ &= \frac{\left(\prod_{d\in\mathcal{D}_{\text{obs},n}} \text{N}\left(x_{n,d} \mid \mu_d(\mathbf{z}_n), \sigma_d^2(\mathbf{z}_n)\right)\right)^{1/\beta} \text{N}(\mathbf{z}_n^{(s)} \mid \mathbf{0}, \mathbf{I})}{q_{\boldsymbol{\phi}}(\mathbf{z}_n^{(s)}|\mathbf{x}_{\text{mis},n}^{(0)}, \mathbf{x}_{\text{obs},n})},\end{aligned}$$

where $\mathcal{D}_{\text{obs},n} \subseteq \{1, \ldots, D\}$ contains the indices of the observed features for the nth data point.

A.6 Supplemental Figures

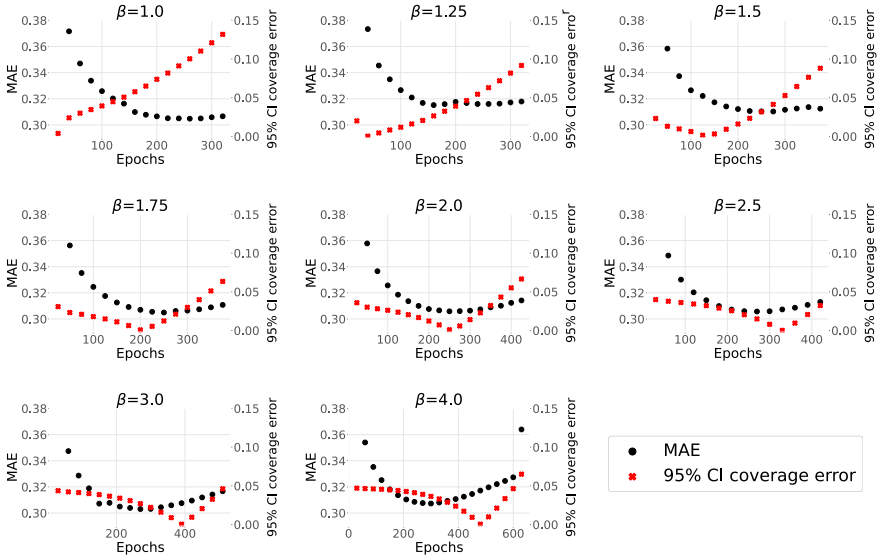


Supplementary Figure A.1: As we increase the number of imputations, M, per missing data point, the mean of the M imputed values gets closer to the true value. The results presented are for the MCAR LGG/GBM dataset.

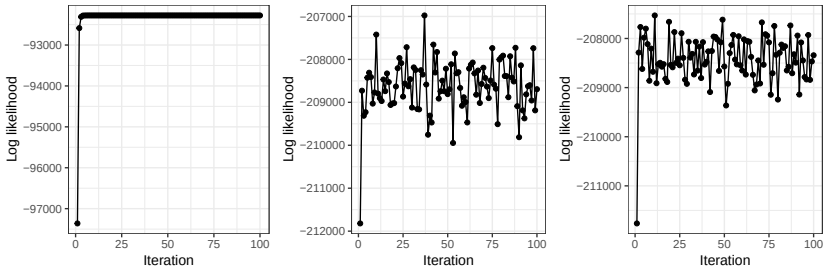


(a) Single imputation (b) PG (c) MWG (d) SIR

Supplementary Figure A.2: **Imputed values across all imputation methods.** Here we report the (multiple) values imputed for the missing data, ranked by their true values (highlighted in red) for **(a)** single imputation (SI) and multiple imputation by **(b)** PG, **(c)** MWG and **(d)** SIR.

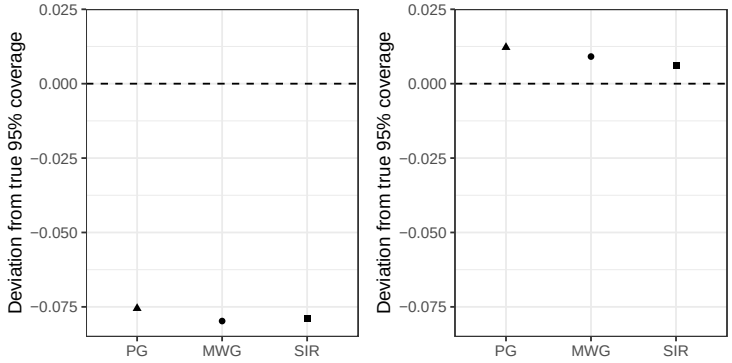


Supplementary Figure A.3: **Results from 5-fold cross-validation to determine optimal model and hyper-parameters.** MAE (black) and EC (red) of 95% CI (computed based on quantiles) at different training epochs. We aim to minimize both of these metrics for optimal training parameters.



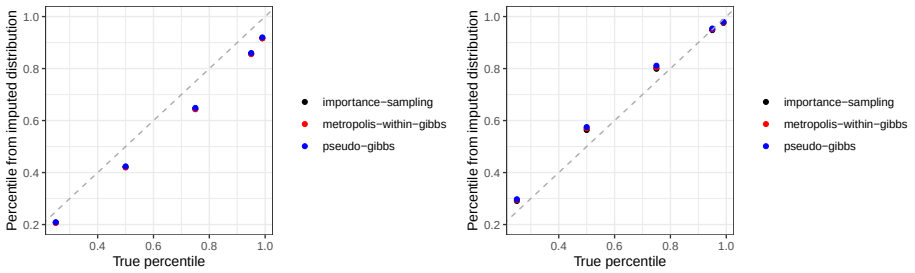
(a) SI (b) PG (c) MWG

Supplementary Figure A.4: **Trace plots monitoring convergence of Markov Chain Monte Carlo schemes for $\beta = 2$ case.** Here we report the log likelihood of the data under the generative model at each iteration for **(a)** single imputation (SI), **(b)** pseudo-Gibbs (PG) and **(c)** Metropolis-within-Gibbs (MWG). For visualization purposes, we show iterations 1 to 100, but ran to 1000 iterations in implementation.



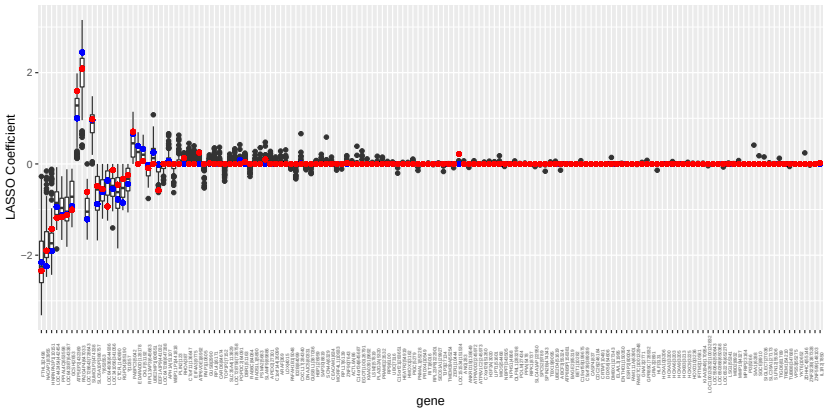
(a) $\beta = 1$ (b) $\beta = 2$

Supplementary Figure A.5: **Deviation from true coverage for all three multiple imputation approaches.** Here we report the deviation from the desired coverage of 95% (showed here by the dotted line) for all three multiple imputation approaches pseudo-Gibbs (PG), Metropolis-within-Gibbs (MWG) and sampling importance resampling (SIR) for **(a)** $\beta = 1$, and **(b)** $\beta = 2$.

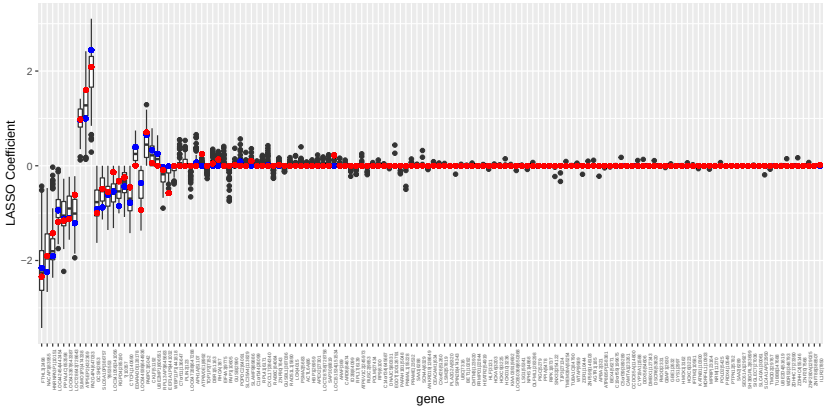


(a) $\beta = 1$ (b) $\beta = 2$

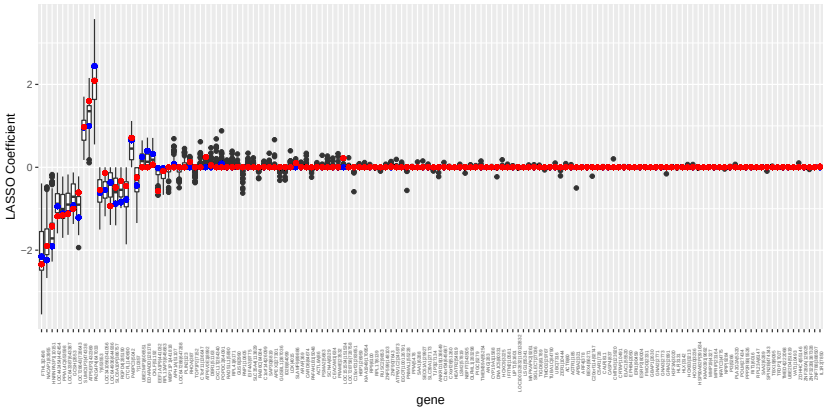
Supplementary Figure A.6: **Coverage by percentiles of all three multiple imputation approaches.** Here we report the coverage evaluated by percentiles across all imputed datasets compared to the true percentiles of 0.25, 0.5, 0.75, 0.95 and 0.99 for all three multiple imputation approaches pseudo-Gibbs (PG), Metropolis-within-Gibbs (MWG) and sampling importance resampling (SIR) for **(a)** $\beta = 1$, and **(b)** $\beta = 2$.



(a) PG

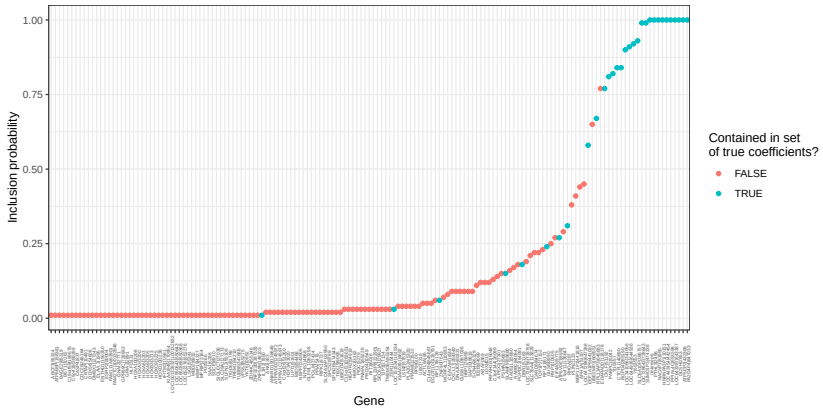


(b) MWG

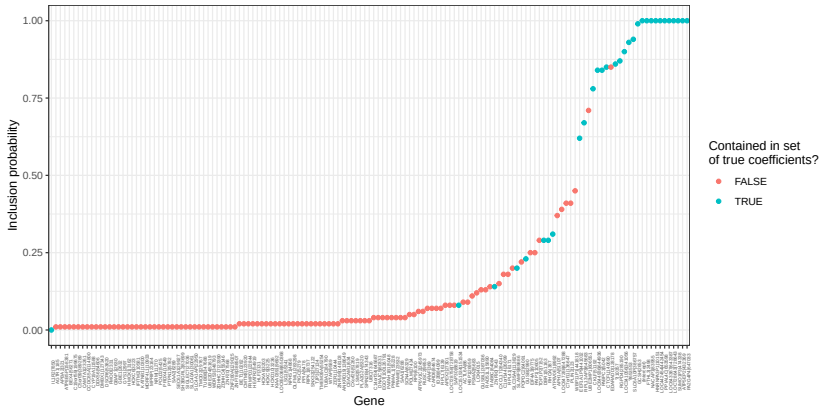


(c) SIR

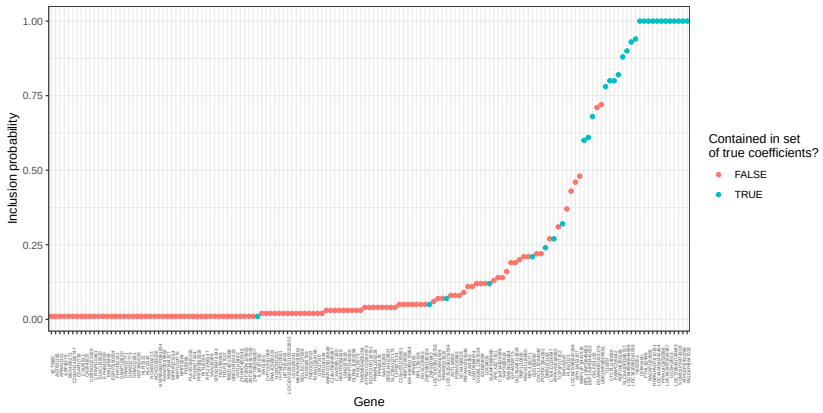
Supplementary Figure A.7: **LASSO regression coefficients across multiple imputation approaches.** Here we report the LASSO regression coefficients across all 100 plausible imputed datasets for **(a)** pseudo-Gibbs (PG), **(b)** Metropolis-within-Gibbs (MWG) and **(c)** sampling importance resampling (SIR). LASSO regression coefficient value for the true dataset is highlighted in red, and for single imputation is highlighted in blue. For the purpose of visualization, the intercept was removed from this plot.



(a) PG

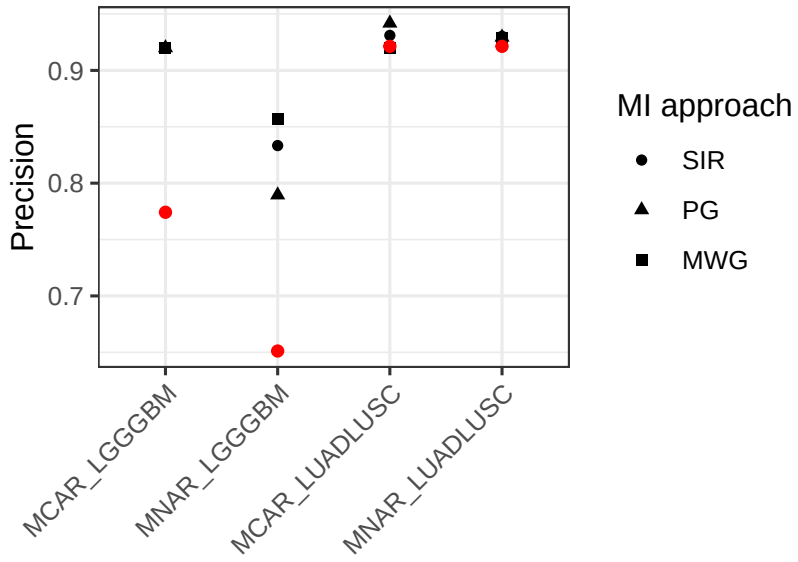


(b) MWG



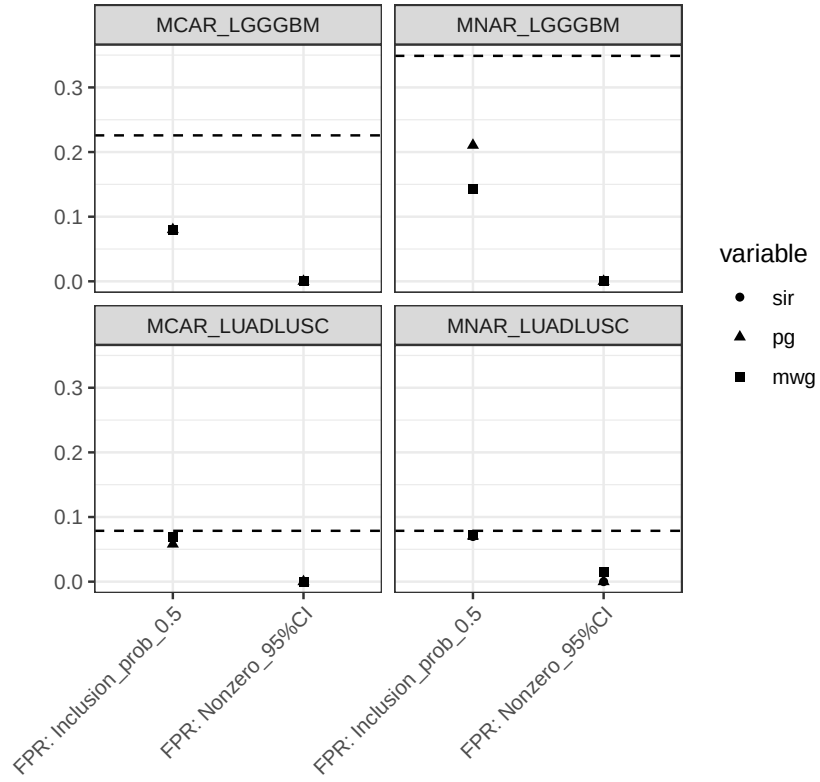
(c) SIR

Supplementary Figure A.8: **Inclusion probability across all non-zero LASSO coefficients across multiple imputation approaches.** Here we report the inclusion probability P_{incl} for all non-zero LASSO regression coefficients for **(a)** pseudo-Gibbs (PG), **(b)** Metropolis-within-Gibbs (MWG) and **(c)** sampling importance resampling (SIR). Genes in this set that are included in the true LASSO coefficients are highlighted in blue, and those not included in the true discriminating gene set are highlighted in pink.



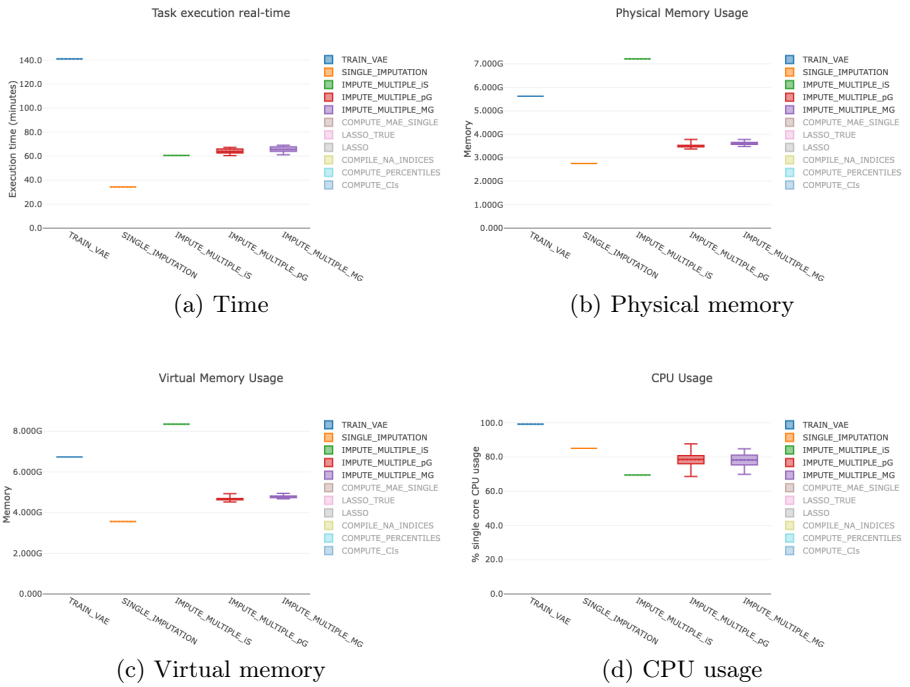
(a) Precision

Supplementary Figure A.9: **PPV for discriminating gene set $P_{\mathbf{incl}} > 0.5$ identified by LASSO across MI strategies compared to SI.** Here we show across all four missing contexts, the precision for the set of genes for $P_{\text{incl}} > 0.5$ across missing scenarios **(a)** MCAR for LGG versus GBM, **(b)** MNAR for LGG versus GBM, **(c)** MCAR for LUAD versus LUSC and **(d)** MNAR for LUAD versus LUSC. With the FPR for single imputation marked with the dotted horizontal line.



(a) FPR

Supplementary Figure A.10: **FPR for both discriminating gene sets identified by LASSO across MI strategies.** Here we show across all four missing contexts, the FPR for the set of genes for (1) $P_{\text{incl}} > 0.5$ and (2) genes across all 100 imputed datasets that do not have zero in the 95% CI for **(a)** MCAR for LGG versus GBM, **(b)** MNAR for LGG versus GBM, **(c)** MCAR for LUAD versus LUSC and **(d)** MNAR for LUAD versus LUSC. With the FPR for single imputation marked with the dotted horizontal line.



(a) Time (b) Physical memory

(c) Virtual memory (d) CPU usage

Supplementary Figure A.11: **Computational burden for MI compared to SI.**

References

1. Mattei, P.A., Frellsen, J.: Miwae: Deep generative modelling and imputation of incomplete data sets. In: International Conference on Machine Learning. pp. 4413–4423. PMLR (2019)