# MultImp: Multiomics Generative Models for Data Imputation

**Ji-Eun Park** [*1] **Wancen Mu** [*1] **Yining Jiao** [*2] **Michael I. Love** [3 1] **Marc Niethammer** [2] **Natalie Stanley** [2]

## Abstract

In biomedical applications, patients are often profiled with multiple technologies or assays to produce a multiomics or multiview biological dataset. A challenge in collecting these datasets is that there are often entire views or individual features missing, which can significantly limit the accuracy of downstream tasks, such as, predicting a patient phenotype. Here, we propose a multiview based deep generative adversarial data imputation model (`MultImp`). `MultImp` improves imputation quality and disease subtype classification accuracy in comparison to several baseline methods across two multiomics datasets. `MultImp` is now publicly available at https://github.com/multimp/multimp.

## 1. Introduction

Across a range of biomedical and clinical applications, it is becoming increasingly common to profile a set of patients with multiple modalities to obtain a so-called 'multiomics' or multiview dataset. In these datasets, each view corresponds to a set of features measured per patient with a particular technology or assay and is intended to summarize a particular biological process. In diverse clinical applications, such as in predicting pregnancy complications (Stelzer et al., 2021; Ghaemi et al., 2018), and subtyping cancers (Vasaikar et al., 2018), the successful integration of multiple biological views has proven to facilitate more accurate patient outcome prediction than through any one view alone.

As multimodal profiling is becoming increasingly prominent in biomedical applications, there are several practical computational challenges that need to be addressed in order to be able to jointly use multiple modalities for downstream unsupervised and supervised tasks. The first problem that has received significant attention focuses on learning a representation for each patient that is consistent across views (Pierre-Jean et al., 2019; Ding et al., 2019). A second class of methods has focused on linking multiview biological data to downstream prediction tasks, such as disease subtype classification (Couture et al., 2019) or predicting continuous outcomes, such as, gestational age in pregnancy (Ghaemi et al., 2018).

A practical limitation in multiview biological datasets is that for various reasons, entire modalities or individual features within a view can be missing in certain patients (Argelaguet et al., 2018). Some practical reasons for missing data include technical errors, patient complications, or cost constraints. Given that each view represents a unique biological process, but that regulatory mechanisms and cross-talk exists (Stelzer et al., 2021; Morton et al., 2019) between features across modalities, it is reasonable to assume that imputation quality for missing features *within* a view can be improved by leveraging insight from *other* views.

Common practice for imputation within a single view is to use the mean feature values across all patients, or across the $k$-nearest patients, or to use a matrix factorization-based approach (Song et al., 2020). More sophisticated methods that target imputing multiomic datasets were developed, for example, based on neural networks, ensemble learning (Song et al., 2020). In multimodal datasets for non-biological applications, generative modeling approaches with input across all views have shown promise in improving imputation quality (Shang et al., 2017). In this work, our primary objective is to apply state-of-the-art generative modeling approaches for imputation in biomedical applications and to systematically study the extent to which it is helpful to perform imputation using information across all views for same patients. Additionally, our presented method further expands on an existing generative modeling approach for this task (Zhang et al., 2020) by adapting it to collectively handle categorical and continuous features, which are common in biomedical datasets.

Here we introduce `MultImp`, a multiview generative modeling approach for imputing multimodal data. Our contributions are as follows.

- `MultImp` adapts CPM-Nets (Zhang et al., 2020) to handle both categorical and continuous features, which is common in biomedical datasets.

[*]Equal contribution [1]Department of Biostatistics, University of North Carolina, Chapel Hill [2]Department of Computer Science, University of North Carolina, Chapel Hill [3]Department of Genetics, University of North Carolina, Chapel Hill. Correspondence to: Michael I. Love <milove@email.unc.edu>, Marc Niethammer <mn@cs.unc.edu>, Natalie Stanley <natalies@cs.unc.edu>.

The 2021 ICML Workshop on Computational Biology. 

- MultImp handles more diverse types of missingness, including random missingness features and view missingness, while CPM-Nets can only handle missing views.
- We evaluate MultImp on two multiview biomedical datasets by measuring imputation accuracy and performance in two disease subtype classification tasks.

## 2. Methods

We define $\mathcal{X}$ as a multiview dataset over $N$ samples. Here, $\mathbf{X}_n = \{\mathbf{x}_n^{(v)}\}_{v=1}^V$ encodes all $V$ dataset views in the $n$th sample. For a given $\mathbf{x}_n^{(v)}$, we define $\mathbf{x}_n^{(v(r))}$ as the vector of continuous features, and $\mathbf{x}_n^{(v(c))}$ as the one-hot version of categorical features. We further represent the number of features in the $v$th view as $I_v$. The set $\mathcal{S}$ denotes a collection of binary indicators that record which values are observed in $\mathcal{X}$. For the $n$th sample, $\mathbf{S}_n = \{\mathbf{s}_n^{(v)}\}_{v=1}^V$ particularly represents the set of these indicators in the $n$th sample across all views. Here, $s_{nj}^{(v)} = 1$ if the $j$th feature is observed in the $v$th view of the $n$th sample, and $s_{nj}^{(v)} = 0$, otherwise.

### 2.1. Model

Our proposed MultImp is a combination of deep generative models and multiview learning for data imputation. Beyond CPM-Nets, we adapt our method to handle both categorical and continuous features and more diverse kinds of missingness than only the view missing case explored in Ref. (Zhang et al., 2020). The overview of our MultImp method is illustrated as Figure 1.

**Multiview Learning** As shown in Figure 1, we apply multiview learning to embed all the samples with arbitrary missingness patterns (missing completely at random, or 'MCAR') into a comprehensive shared latent space. Here, the latent representation of samples is consistent acoss views and each view is able to provide distinctive information to better encode the samples in the latent space. We evaluate the quality of the latent representations and the generators using the reconstruction loss, $L_{rec}$, of observed features as equation (1).

**Categorical and Continuous Features** In order to recover missing categorical features, different losses are imposed on categorical and continuous features, respectively. Here, $G_{v(r)j}(\mathbf{h}_n)$ represents the reconstructed $j$th continuous feature in the $v$th view from the generators, and $\mathbf{h}_n$ is the latent representation for $n$th sample. Similarly, we let $G_{v(c)j}(\mathbf{h}_n)$ be the the one-hot version of the reconstructed $j$th categorical feature in the $v$th view. We define $\mathcal{L}_{r(c)}$ and $\mathcal{L}_{r(r)}$ as the cross entropy (CE) and L2 loss for recovering categorical and continuous features, respectively, in equations (2) and (3).

$$\mathcal{L}_{rec} = \mathcal{L}_{r(r)} + \mathcal{L}_{r(c)} \tag{1}$$

$$\mathcal{L}_{r(r)} = \Sigma_{n,v}\Sigma_{j=1}^{I_{v(r)}} s_{nj}^{(v(r))} ||G_{v(r)j}(\mathbf{h}_n) - x_{nj}^{(v(r))}||^2 \tag{2}$$

$$\mathcal{L}_{r(c)} = \Sigma_{n,v}\Sigma_{j=1}^{I_{v(c)}} s_{nj}^{(v(c))} (-\mathbf{x}_{nj}^{(v(c))} \cdot \log(G_{v(c)j}(\mathbf{h}_n))) \tag{3}$$

Here, $\cdot$ is the notation for inner product.

**Generative Adversarial Networks (GANs)** GANs (Goodfellow et al., 2014) have shown remarkable potential for data imputation. MultImp consists of two sets of GAN modules: a set of generators $G(\cdot)$ to learn the distribution $p(\mathbf{x}_{data}^{(v)})$ over the data $\mathbf{x}$ in the $v$th view from $\mathbf{h}_n$, and a corresponding set of discriminators $D(\cdot)$ to discriminate real samples from generated samples. These two sets of modules are trained in an adversarial manner. Considering the random missingness in real samples, we define our GAN loss in equation 4 as,

$$\begin{aligned}\mathcal{L}_{adv} = \Sigma_{n=1}^N \Sigma_{v=1}^V \Sigma_{m=1}^M [&\log D_v(\mathbf{s}_m^{(v)} \circ \mathbf{x}_m^{(v)}) \\ &+ \log(1 - D_v(\mathbf{s}_n^{(v)} \circ G_v(\mathbf{h}_n)))]\end{aligned} \tag{4}$$

Here, $\circ$ is the notation for Hadamard product.

**Overall Loss** The overall loss, $\mathcal{L}$ (equation 5), is composed of the view reconstruction loss and the GAN loss. View reconstruction loss is for precisely recovering missing values and the GAN loss ensures that our imputation is realistic and can retain significant information and variance from the original data. The loss, $\mathcal{L}$, is defined as,

$$\mathcal{L} = \min_G \max_D \min_{\mathbf{h}} \mathcal{L}_{adv} + \mathcal{L}_{rec}. \tag{5}$$

## 3. Experiments

In each dataset, we sequentially excluded varying percentages of matrix elements at random, with the rate of exclusion ranging between 10% and 50%. After imputing back the excluded (missing) features, we calculated 1) the average root mean squared error (RMSE) for imputed continuous features, 2) the average accuracy (Jaccard Similarity) for categorical features, and 3) the average accuracy in a disease subtype classification task. Each experiment was repeated five times.

### 3.1. Datasets

In our experiments, we used the following two multiview biomedical datasets.

- **ADNI**. Alzheimer's Disease Neuroimaging Initiative (ADNI) data (Mueller et al., 2005) is a dataset involving multiple types of experimental assays for understanding Alzheimer's Disease. We used two views in

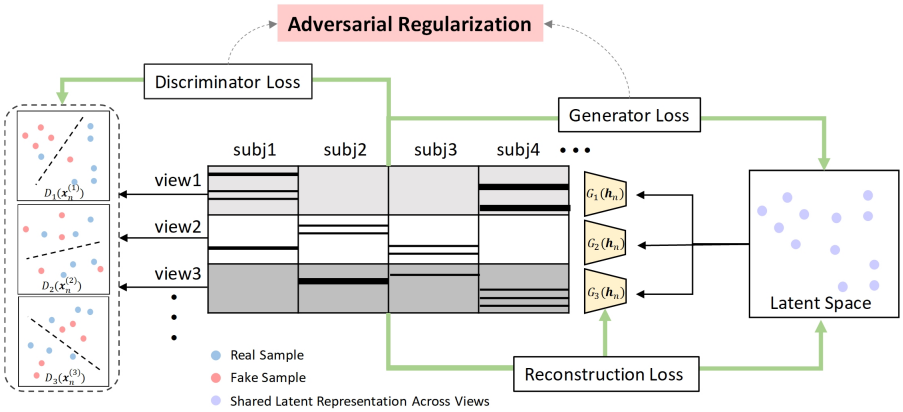


*Figure 1.* Illustration of *MultImp*. Each sample is represented by a shared $S$-dimensional latent representation, $\mathbf{h}_n \in R^S$. Under different generators, $\{G_v(\mathbf{h}_n)\}_{v=1}^V$, the features across different views can be recovered from their latent representation $\mathbf{h}_n$. View reconstruction loss, $\mathcal{L}_{rec}$, is used for optimizing the generators, $\{G_v(\mathbf{h}_n)\}_{v=1}^V$, and the latent representation, $\mathbf{h}_n$, iteratively. The discriminators, $\{D_v(\mathbf{x}^{(v)})\}_{v=1}^V$ determine whether a sample was reconstructed by a generator or is a true original one. The discriminators $\{D_v(\mathbf{x}^{(v)})\}_{v=1}^V$ specify a criterion to determine whether the reconstructed data from the latent space is sufficiently realistic.

this dataset that record 1) clinical information and 2) gene expression profiling. Across views, there are 744 total matched samples.

- **TCGA-Glioma**. The TCGA Glioma dataset (Vasaikar S., 2017) contains seven views to describe Glioma. These views include, 1) clinical annotation, 2) somatic mutation data (gene level SNVs), 3) RNAseq (normalized counts via the Illumina HiSeq platform, Gene-level, Normalized log2 RPKM), 4) copy number variation (focal level, GISTIC2 log ratio), 5) copy number variation (gene level, GISTIC2 log ratio), 6) miRNA expression for tumor samples (Normalized, RPM), and 7) methylation data of tumor samples at Gene level (Beta values, Illumina HM450K platform). Irrelevant (no variance across patients) genes from the RNAseq and mutation views were removed. Only 425 complete samples were kept for model evaluation.

*Table 1.* MultImp Variations

| Variations | GAN | Multiview | CE Loss |
|---|---|---|---|
| (a) | √ | √ | √ |
| (b) | √ | × | √ |
| (c) | × | √ | √ |
| (d) | × | × | √ |
| (e) | √ | √ | × |
| (f) | √ | × | × |
| (g) | × | √ | × |
| (h) | × | × | × |

### 3.2. MultImp Variations + Baseline Definitions

We compare MultImp to three baselines available in the Python package 'FancyImpute' (Alex Rubinsteyn and Sergey Feldman). These baselines include Mean Imputation, KNN Imputation (Crookston & Finley, 2008) and Matrix Factorization Imputation (Koren et al., 2009).

To investigate the benefit from using 1) GANs, 2) multiview learning, and 3) the cross entropy (CE) loss for categorical features, different variants of MultImp were implemented. We define 8 variants of MultImp in Table 1. Here '×' in the multiview column implies all views were concatenated into a single view.

### 3.3. Patient Subtype Classification Tasks

To evaluate whether the imputation approach can be used in downstream classification tasks, we formulated disease subtype classification tasks for each of the two datasets. Since the number of features is dramatically larger than the number of samples, we first concatenated all the features together (original and imputed) and applied principal component analysis (PCA) to reduce the dataset dimension to 128. A support vector machine (SVM) was then trained to predict the disease subtype based on the 128 features defined by PCA. In the ADNI dataset we formulated a multiclass classification problem to predict the 4 disease status (AZ, NC, MCI, LMCI). In the TCGA Glioma dataset, we aimed to predict astrocytoma from other subtypes of glioma.

### 3.4. Results

In both the ADNI (Fig. 2(a), 2(b), 2(c)) and TCGA Giloma datasets, (Fig. 2(d), 2(e), 2(f)), MultImp or its variants outperform all three baselines on all evaluation tasks (Fig. 2). MultImp also imputes categorical features quite robustly across multiple rates of missingness. Specifically, in Figs. 2(b) and 2(e), the accuracy in imputing categorical features achieved by baseline methods have comparably large variance while the variance fluctuated only between 0.82 to 0.85 in the MultImp variants. In Fig. 2(a) and 2(d), we show that GANs improve the imputation quality for continuous

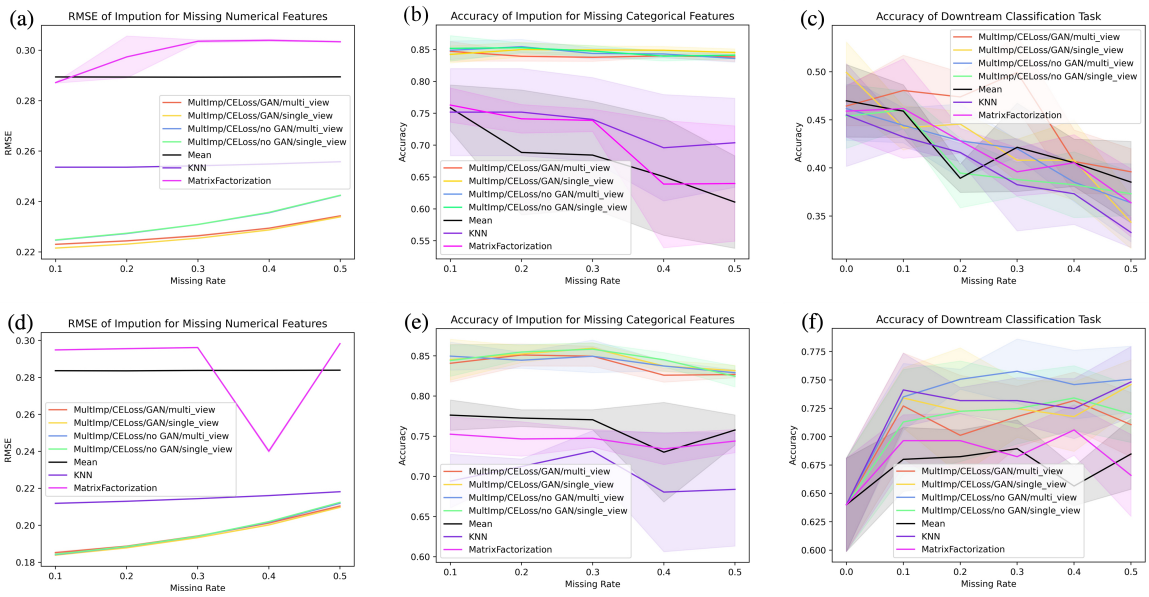


*Figure 2.* Imputation Quality: (a) ADNI imputation RMSE, (b) ADNI imputation accuracy for categorical features, (c) ADNI disease subtype classification accuracy, (d) TCGA imputation RMSE, (e) TCGA imputation accuracy for categorical features, (f) TCGA disease subtype classification accuracy

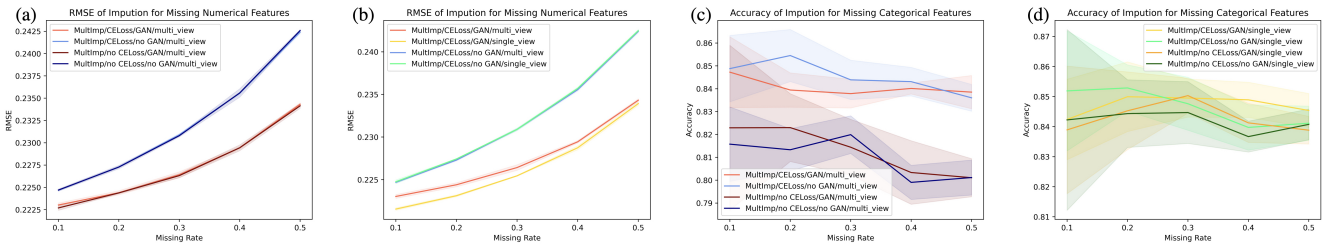


*Figure 3.* Imputation Quality Difference By Adding Classification Loss for Categorical Features: (a) ADNI multiview RMSE, (b) ADNI single-view RMSE, (c) ADNI multiview categorical accuracy, (d) ADNI single-view categorical accuracy

features but not necessarily for categorical features. Further, multiview learning was not as helpful as GANs for imputing categorical and continuous features. Alternatively, in the disease-subtype classification tasks, we observed that multiview learning usually yields the best performance (Fig. 2(c), 2(f)).

As shown in Figs. 3(a) and 3(b), there is little benefit from CE loss for imputing continuous features, likely because there are very few categorical features compared to continuous features. However, in Figs. 3(c) and 3(d), we show that adding the cross entropy loss helped to improve the categorical feature imputation accuracy.

Fig. 2(f) shows that using imputed data can more accurately predict disease subtypes than using the original data. This is potentially due to the fact that the imputation with `MultImp` can capture the distinguishable information for recognizing glioma subtypes, while reducing noise across views.

## 4. Discussion

Our experiments show that for both multiview and single-view learning, GANs are beneficial for imputation. The benefit of multiview learning is more apparent in the disease subtype classification tasks but less noticeable with respect to imputation precision. This observation does not align with previous studies on non-biological datasets (Zhang et al., 2020). We suspect that in biological datasets, the assumption that each view has the same number of representatives in the shared latent space is unreasonable, and it is therefore of interest in future work to consider how to allow for more flexible shared view-specific latent representations.

Overall, we show that adding a cross-entropy loss does not significantly improve the imputation quality of numerical features. We expect this modification to be more beneficial in datasets with a higher proportion of categorical features. Further exploration is needed to understand the implications of missingness in multiomics datasets, especially with respect to predicting clinical outcomes (Lim et al., 2021).

## References


Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python. URL https://github.com/iskandr/fancyimpute.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6), 2018.

Couture, H. D., Kwitt, R., Marron, J., Troester, M., Perou, C. M., and Niethammer, M. Deep multi-view learning via task-optimal cca. *arXiv preprint arXiv:1907.07739*, 2019.

Crookston, N. L. and Finley, A. O. yaimpute: an r package for knn imputation. *Journal of Statistical Software. 23 (10). 16 p.*, 2008.

Ding, H., Sharpnack, M., Wang, C., Huang, K., and Machiraju, R. Integrative cancer patient stratification via subspace merging. *Bioinformatics*, 35(10):1653–1659, 2019.

Ghaemi, M. S., DiGiulio, D. B., Contrepois, K., Callahan, B., Ngo, T. T., Lee-McMullen, B., Lehallier, B., Robaczewska, A., Mcilwain, D., Rosenberg-Hasson, Y., et al. Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics*, 35(1): 95–103, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37, 2009.

Lim, D. K., Rashid, N. U., Oliva, J. B., and Ibrahim, J. G. Handling non-ignorably missing features in electronic health records data using importance-weighted autoencoders, 2021.

Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., et al. Learning representations of microbe–metabolite interactions. *Nature methods*, 16(12):1306–1314, 2019.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

Pierre-Jean, M., Deleuze, J.-F., Le Floch, E., and Mauger, F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics*, 2019.

Shang, C., Palmer, A., Sun, J., Chen, K.-S., Lu, J., and Bi, J. Vigan: Missing view imputation with generative adversarial networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 766–775. IEEE, 2017.

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., and Deng, H.-W. A review of integrative imputation for multi-omics datasets. *Frontiers in Genetics*, 11:1215, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.570255. URL https://www.frontiersin.org/article/10.3389/fgene.2020.570255.

Stelzer, I. A., Ghaemi, M. S., Han, X., Ando, K., Hédou, J. J., Feyaerts, D., Peterson, L. S., Rumer, K. K., Tsai, E. S., Ganio, E. A., et al. Integrated trajectories of the maternal metabolome, proteome, and immunome predict labor onset. *Science Translational Medicine*, 13(592), 2021.

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic acids research*, 46(D1):D956–D963, 2018.

Vasaikar S., Straub P., W. J. Z. B. Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, 46:D956–D963, 2017. doi: https://doi.org/10.1093/nar/gkx1090.

Zhang, C., Cui, Y., Han, Z., Zhou, J. T., Fu, H., and Hu, Q. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020.