# Comparative Assessment and Novel Strategy on Methods for Imputing Proteomics Data

**Minjie Shen**
Virginia Polytechnic Institute and State University
**Yi-Tan Chang**
Virginia Polytechnic Institute and State University
**Chiung-Ting Wu**
Virginia Polytechnic Institute and State University
**Sarah J. Parker**
Cedars Sinai Medical Center
**Georgia Saylor**
Wake Forest University
**Yizhi Wang**
Virginia Polytechnic Institute and State University
**Guoqiang Yu**
Virginia Polytechnic Institute and State University
**Jennifer E. Van Eyk**
Cedars Sinai Medical Center
**Robert Clarke**
University of Minnesota
**David M. Herrington**
Wake Forest University
**Yue Wang** ( yuewang@vt.edu )
Virginia Polytechnic Institute and State University

**Research Article**

**Keywords:** Comparative assessment, novel strategy, methods, imputing proteomics data, major issue, quantitative proteomics analysis,

**Posted Date:** August 31st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-841815/v1

**Version of Record:** A version of this preprint was published at Scientific Reports on January 20th, 2022. See the published version at https://doi.org/10.1038/s41598-022-04938-0.

# Comparative assessment and novel strategy on methods for imputing proteomics data

Minjie Shen[1,#], Yi-Tan Chang[1,#], Chiung-Ting Wu[1,#], Sarah J. Parker[2], Georgia Saylor[3], Yizhi Wang[1], Guoqiang Yu[1], Jennifer E. Van Eyk[2], Robert Clarke[4], David M. Herrington[3], and Yue Wang[1,†]

[#]Equal contribution

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [2]Advanced Clinical Biosystems Research Institute, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA; [3]Department of Internal Medicine, Wake Forest University, Winston-Salem, NC 27157, USA; [4]The Hormel Institute, University of Minnesota, Austin, MN 55912, USA

Running head: Proteomics missing value imputation

[†]Correspondence: Yue Wang
Dept. of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
900 N. Glebe Road
Arlington, VA 22203, USA
Email: yuewang@vt.edu

# Abstract

Missing values are a major issue in quantitative proteomics analysis. While many methods have been developed for imputing missing values in high-throughput proteomics data, a comparative assessment of imputation accuracy remains inconclusive, mainly because mechanisms contributing to true missing values are complex and existing evaluation methodologies are imperfect. Moreover, few studies have provided an outlook of future methodological development. We first re-evaluate the performance of eight representative methods targeting three typical missing mechanisms. These methods are compared on both simulated and masked missing values embedded within real proteomics datasets, and performance is evaluated using three quantitative measures. We then introduce fused regularization matrix factorization, a low-rank global matrix factorization framework, capable of integrating local similarity derived from additional data types. We also explore a biologically-inspired latent variable modeling strategy - convex analysis of mixtures - for missing value imputation and present preliminary experimental results. While some winners emerged from our comparative assessment, the evaluation is intrinsically imperfect because performance is evaluated indirectly on artificial missing or masked values not authentic missing values. Nevertheless, we show that our fused regularization matrix factorization provides a novel incorporation of external and local information, and the exploratory implementation of convex analysis of mixtures presents a biologically plausible new approach.


# Introduction

Liquid chromatography coupled to mass spectrometry (LC-MS) is a popular method for high-throughput identification and quantification of thousands of proteins in a single analysis [1,2]. The LC-MS signals can be displayed in a three-dimensional space consisting of the mass-to-charge ratios, retention times and intensities for the observed peptides. However, this approach suffers from many missing values at the peptide or protein level, which significantly reduces the amount of quantifiable proteins with an average of 44% missing values from traditional LC-MS workflows [3-5].

While there are multiple causes for this missingness, three typical missing mechanisms are widely acknowledged. Low abundant proteins may be missing because their concentration is below the lower limit of detection (LLD); while poorly ionizing peptides or

problems in technical pre-processing may cause proteins to be missing not at random (MNAR) [6]. However, missingness may also extend to mid- and even high-range intensities [5], statistically categorized into missing at random (MAR) and missing completely at random (MCAR) [7]. MAR is actually missing conditionally at random given the observed data distribution or underlying parametric covariates. MCAR depends on neither observed nor missing data, thus the incomplete data are representative of the entire dataset. While MAR allows prediction of the missing values based on observed data, unfortunately, the MAR and MNAR conditions cannot be distinguished based on the observed data because by definition missing values are unknown [8,9]. More importantly, missing values in reality can originate from a mix of both known and unknown missing mechanisms [7,10].

A common solution for missingness is to impute the missing values based on assumed missing mechanisms. However, this approach can introduce a profound change in the distribution of protein-level intensities because most methods are only designed for a single missing mechanism. These changes can have unpredictable effects on downstream differential analyses. While many imputation methods have been adopted for imputing missing values in proteomics data, comparative evaluation of their relative performance remains inconclusive. Moreover, few studies provide an outlook on how best to address unresolved problems or future development directions [4,9,10].

To better understand the strengths and limitations of both imputation methods and assessment designs, we conduct a collective assessment of eight representative methods involving three typical missing value mechanisms in conjunction with authentic missing values. Compared using a set of realistic (preserving data distribution) simulations derived from real proteomics data sets, the performance of the selected methods is measured by three criteria [11], root-mean-square error (RMSE), normalized root-mean-square error (NRMSE), and sum of ranks (SOR). Several important observations are evident from this comparison study. First, while imputation methods perform differentially under various missing mechanisms, algorithmic parameter settings, and preprocessing procedures, some methods consistently perform better than others across a range of realistic simulation studies. Second, the quality of performance assessment depends on the efficacy of simulation designs; a more realistic simulation design should include authentic missing values and preserve the original overall data distribution. Third, existing assessment methodologies are imperfect in that performance is indirectly assessed on imputing either artificial or masked, but not authentic missing values (see Discussion section).

To explore a more integrative strategy for improving imputation performance, we discuss a low-rank matrix factorization framework with fused regularization on both sparsity and similarity – Fused Regularization Matrix factorization (FRMF) [12-14], which can naturally integrate other-omics data such as gene expression or clinical variables. We also introduce a biologically-inspired latent variable modeling strategy - Convex Analysis of Mixtures (CAM) [14,15], which explicitly formulates a data matrix as mixtures of underlying biological archetypes and performs missing value imputation on original intensity data (before log-transformation). Preliminary results on real proteomics data are provided together with an outlook into future development directions.

# Results

### Experimental design and protocol

We selected eight representative methods for comparative assessment, based on their intended missing mechanism(s) and imputation principles (summarized in **Figure 1**). One method (Min/2) is devoted to MNAR (LLD) [7], two methods (swKNN and pwKNN) are tailored to MAR (local-similarity) [16], and five methods (Mean, PPCA, NIPALS, SVD, and SVT) are intended for MCAR/MAR (global-structure or low-rank matrix factorization) [7,10,17-19]. We then explored and tested several variants of FRMF and CAM, where local similarity information is obtained from baseline or other data acquired from the same samples.

We conducted the comparative assessments in two complementary simulation settings. In simulation setting 1, the simulation data were generated from the observed data portion (no authentic missing value) of a real proteomics dataset, where artificial missing values were introduced by two typical missing mechanisms and used for performance assessment. In simulation setting 2, the simulation data were generated from the complete data matrix (including authentic missing values) of a real proteomics dataset, where a small percentage of data points were randomly set-aside (masked values) and used solely for performance assessment. Preprocessing eliminates proteins with missing rates higher than 80% and then performs log2 transformation [20]. Parameters were optimized for each imputation method by parameter sweeping over a wide range of settings at each missing rate. The overall experimental workflow is given in **Figure 2**.

### Real proteomics data

Real LC-MS proteomics data form the base from which the simulation data sets were produced [6]. Data were acquired using label-free data-independent acquisition (DIA) protocol,

and protein level output was generated by mapDIA [21]. The resulting dataset contains 200 samples associated with 2,682 proteins measured in human left anterior descending (LAD) coronary arteries collected as part of a study of coronary and aortic atherosclerosis [22]. Data were produced in three separate batches, indexed as A, B, and C; and all data passed the quality control and preprocessing procedures [6,22], as summarized in **Table 1** (Supplementary Information). To avoid unknown and unnecessary batch effects, all simulation datasets were generated from batch A dataset that has the largest sample size (n=98). For the simulations on other batches, please see the Supplementary Information.

**Table 1.** Summary of real proteomics datasets used in this work.

| | Sample size | Protein size | Total Missing Rate #MV/(#Sample*#Protein) | Setting #1 protein size (non-missing proteins) | Setting #2 protein size (proteins with ≤ 80% missing rate) |
|---|---|---|---|---|---|
| Batch A | 98 | 2107 | 24.67% | 751 (35.64%) | 1935 (91.84%) |
| Batch B | 55 | 2604 | 29.63% | 819 (31.45%) | 2324 (89.25%) |
| Batch C | 47 | 2590 | 25.52% | 976 (37.68%) | 2325 (89.77%) |

### Simulation data generated from the observed portion of the data matrix (Setting 1)

Based on the observed portion of data matrix (without authentic missing values), we adopted a hybrid missing data model and used the R package imputeLCMD to introduce artificial missing values while preserving the original observed data patterns [23]. Specifically, MCAR missing values were introduced by randomly replacing some data points with 'NA' (not available) according to the designed missing rates (approximately from 1% to 50%); MNAR missing values were introduced by quantile cut-off for the full dataset [7,10,20]; and mixed MCAR and MNAR missing values were introduced by assigning $(1-\beta)$ portion of MCAR and $\beta$ portion of MNAR; corresponding to missing rate $\alpha$ and $\beta = 0, 0.1, 1$ (Supplementary Information).

### Simulation data with set-aside masked values from the full data matrix (Setting 2)

In this simulation setting, we used the full data matrix (including both observed and authentic missing values) from the human coronary proteomics dataset. To preserve the original patterns of both observed and authentic missing values, for each protein, a small percentage of data points in the complete data matrix were randomly set-aside as 'NA' (masked values)

with the masking rate(s) proportional to the authentic missing rate(s). This procedure was repeated for all proteins and the masked values were considered as a mix of MNAR and MAR conditioned on the observed missing rates and data patterns (**Figure 3**, Supplementary Information).

### Performance assessment focused on MNAR (Setting 1)

As shown in **Figure 4** (see additional results in Supplementary Information), under an MNAR missing mechanism assumption, SVT and Min/2 yielded the best performance in both simulation settings, and the relative performance of SVT and Min/2 depends on the missing rates and criterion used for evaluation. It is important to reiterate that in reality the MAR and MNAR conditions cannot be distinguished based on the observed data because by definition missing values are unknown [8,9]. As expected, the MNAR-devoted method, Min/2, performs much better than the others. The baseline method Mean performs worst among all methods (see additional results in Supplementary Information). Note that SOR increases expectedly when the missing rate increases because SOR is positively associated with the number of missing proteins, therefore the value of SOR may not imply the absolute performance of a method.

### Performance assessment focused on MCAR (Setting 1)

Imputation performance of the eight methods on the MCAR mechanism is summarized in **Figure 5** (see additional results in Supplementary Information). Experimental results show that NIPALS performs better than all other methods in both simulation settings and for almost all three evaluation criteria. SVT is the best method when RMSE is used. Min/2 performs the worst among all other methods in all cases, likely due to its design for the MNAR mechanism. Mean performs consistently poorly over different missing rates, with only except for Min/2 showing a worse performance. While all methods perform worse when the total missing rate increases, the ranking of their relative performances remains unchanged (see additional results in Supplementary Information).

### Performance assessment focused on authentic missing values (Setting 2)

As shown in **Figure 6** (see additional results in Supplementary Information), NIPALS, SVT, and protein-wise or sample-wise KNN achieve the best performance, where authentic missing values are dominant and where imputation accuracy is evaluated on the masked values. This observation is consistent with what has been reported previously [5,6,11]. As expected, the MNAR-devoted method Min/2 has the worst performance. Similar to the case of MCAR,

low-rank methods and local-similarity methods perform worse when the total missing rate increases, and among these methods, SVD and PPCA perform worse than the baseline method Mean when the total missing rate is large. Note that because low abundant proteins often have higher authentic missing rates and accordingly higher masking rates, more low abundant proteins (possibly the minimum values) are masked than highly expressed proteins. Thus, the counterintuitive decrease in NRMSE by Min/2 is expected when the authentic missing rate increases.

### Evaluation of the FRMF method focused on authentic missing values (Setting 2)

In this study we aimed to experimentally test whether the FRMF method that integrates local similarity derived from within and/or external data could improve imputation accuracy as compared with global low-rank SVD, an existing approach based on a similar principle. We evaluated three variants of the FRMF method. RMF serves as a baseline sparsity regularized matrix factorization algorithm [13]; FRMF_self introduces a fused-regularization utilizing the similarity among samples embedded within the data matrix; and FRMF_cross_patho exploits external pathological scores using a fused-regularization strategy where the pathological scores are the qualitative percentages of the intimal surface involvement of various atherosclerotic changes graded by pathologists [6].

The experimental results are shown in **Figure 7**. While RMF performs comparably with SVD and starts to perform better when missing rates are larger than 25%, both FRMF_self and FRMF_cross_patho performs significantly better than RMF. This preliminary result implies a potential benefit for combining global low-rank and local-similarly regularizations, and also for leveraging external information via fused regularization.

### Evaluation of the CAM method focused on authentic missing values (Setting 2)

In this study we aimed to experimentally test whether the CAM method that exploits biologically interpretable latent variable models in the matrix factorization could improve imputation accuracy as compared with the top existing approaches that use a similar principle (i.e. SVT; NIPALS). Accordingly, based on biologically-inspired latent variable modeling of complex tissues [14,15], we proposed and evaluated three variants of a CAM based imputation strategy. CAM_complete performs CAM based imputation using the non-missing portion of full data matrix; CAM_SVT and CAM_NIPALS perform CAM based imputation using full data matrix where the input data matrix is initialized by either SVT or NIPALS, respectively.

The experimental results are shown in **Figure 8**. As expected, CAM_complete performs much better than the baseline method Mean. More importantly, both CAM_NIPALS and CAM_SVT consistently performs better than NIPALS and SVT - the two top performers from our earlier comparative assessment. This preliminary result shows that biologically-plausible latent variable modeling may potentially improve imputation accuracy within the framework of low-rank optimization.

## Discussion

The ability to simulate the missing values mechanisms (MNAR, MAR, or MCAR) depends on the efficacy of the tools applied. However, because simulation relies on the statistics estimated from the observed portion of a data matrix, and so the artificial missing values introduced cannot fully resemble authentic missing mechanisms and/or patterns present in the original overall data distribution. More critically, performance can only be assessed on evaluating the imputed artificial not authentic missing values, because authentic missing values are intrinsically unknown and the overall data distribution may be distorted by the introduced artificial missing values. While it may be informative to compare the impact of the imputation versus non-imputation on some subsequent data analysis in the future, we have opted to focus on assessing direct imputation accuracy, because the evaluation using subsequent analysis would be indirect and task-dependent.

To address the aforementioned issues in the presence of authentic missing values, a small percentage of set-aside values were introduced into the complete data matrix and used solely for the purpose of assessment. Because masked values are randomly assigned onto both observed and authentic missing values, the simulation maximally preserves the original overall data distribution. Masked values may represent a mix of MNAR (high missing rate associated with low protein abundance) and MAR (joint distribution of both observed and authentic missing values). However, performance is assessed indirectly on imputing masked not authentic missing values. An interesting addition may be to include the performance accuracy in estimating non-missing values by the imputation function.

Imputation accuracy could be affected by data preprocessing and algorithmic parameter setting. In this study, sample-wise normalization and protein-wise standardization are performed based on the requirements of each method. Data preprocessing affects the scale of NRMSE, but relative performance across various methods remains consistent. While imputation performance varies with parameter setting, there is no theoretical guideline for

optimizing the parameter setting. Nevertheless, a cross-validation approach could be used to optimize algorithms parameters in future work. Here, we used grid search to tune the parameters based on the consideration that when all possible parameters are enumerated the optimal setting can be reached. Moreover, other correlation-based relative performance measures (Pearson and/or Spearman) can be calculated on introduced missing or masked values and non-missing values.

The observation that imputation methods based on low-rank matrix factorization (e.g. SVT or NIPALS) perform consistently better than other ones (in both simulation settings) is consistent with the reports from other similar evaluation studies [4,11]. While the true missing mechanisms in proteomics data are unknown and hard to simulate, the better performance of SVT or NAPALS may be expected particular in the cases with relatively small ample size. Missing value imputation is principally an unsupervised learning task, and the sample size is determined by the number of observed values over both samples and proteins. Concerning unavoidable noise/outlier and sample heterogeneity embedded within the observed values of data matrix, related to possibly much smaller 'effective' sample size, SVT or NIPALS leverages low-rank regularization to avoid potential overfit of the model to noise/outlier.

FRMF is a novel integrated imputation approach with promising preliminary results. However, the effectiveness of FRMF for improving global low-rank matrix completion methods depends on both sample diversity and the complementary nature of additional and relevant measurements such as the complementary roles of local similarity and global structure. For example, FRMF imputation may be performed on combined biologically diverse sample groups, or local similarity derived from gene expressions may be incorporated to impute protein missing values on the same samples.

The CAM method we propose represents a new direction for future work. For example, CAM may be integrated into FRMF to leverage local similarity derived from complementary information on the same samples and the input data matrix for CAM may be initialized by missing-value-insensitive NMF (nonnegative matrix factorization) [12]. More importantly, CAM performs missing value imputation using the original intensities rather than log-transformed data. This approach is both biologically plausible and mathematically more rigorous because log-transformation violates the linear nature of low-rank matrix factorization [24].

In future work, support vector machine or artificial neural network (ANN) based methods may be considered as emerging imputation competitors [11], and a combination approach utilizing an ensemble of strategies could be explored [4]. Furthermore, some

advanced hyperparameter auto-search tools could be adapted to optimize the hyperparameters often embedded within various imputation algorithms [25]. Conclusions and insights from more recent similar evaluation work should also be considerred [11].

# Method

### Brief introduction to the eight existing methods

- **Min/2 (half minimum):** Taking MNAR as the missing mechanism, for each protein the missing values are estimated as half the minimum value of the observed intensities in that protein across all samples [6,9].
- **Mean:** For MAR/MCAR as the missing values mechanism, for each protein we replaced the missing values with the mean value of the observed intensities in that protein across all samples [6,9].
- **swKNN (sample-wise k-nearest neighbors):** Taking MAR as the missing values mechanism, we leveraged local similarity among samples for each protein, replacing the missing values with the weighted average of observed intensities in that protein proportional to the proximities of k-nearest neighboring samples [9].
- **pwKNN (protein-wise k-nearest neighbors):** Where MAR was the presumed missing values mechanism, we leveraged local similarity among proteins for each sample, replacing the missing values with the weighted average of observed intensities in that sample proportional to the proximities of k-nearest neighboring proteins (with protein-wise normalization) [9].
- **PPCA (probabilistic PCA):** For MCAR/MAR as the missing values mechanism, a low-rank probabilistic PCA matrix factorization was estimated by the expectation maximization (EM) algorithm and then used to impute missing values [26].
- **NIPALS (non-linear estimation by iterative partial least squares):** Taking MCAR/MAR as the missing values mechanism, a low-rank missing-data-tolerant PCA matrix factorization was estimated by iterative regression and then used to impute missing values [27,28].
- **SVD (SVDImpute):** For MCAR/MAR as then missing values mechanism, a low-rank SVD matrix factorization was estimated by the EM algorithm and used to impute missing values [27,29].
- **SVT (singular value thresholding):** Where we assumed MCAR/MAR to be the missing values mechanism, a low-rank SVT matrix factorization was estimated by

iteratively solving a nuclear norm minimization problem and then used to impute missing values [19].

### Performance measures

Three quantitative measures were used to evaluate imputation accuracy, namely Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), and Sum of Ranks (SOR). Specifically, RMSE and NRMSE are given by [30,31]

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\Omega}(\hat{X}_{\Omega}-X_{\Omega})^2}{|\Omega|}},\quad \mathrm{NRMSE} = \sqrt{\frac{\sum_{\Omega}(\hat{X}_{\Omega}-X_{\Omega})^2}{|\Omega|\sigma^2_{X_{\Omega}}}},$$

respectively, where $\Omega$ is the index set of missing values in complete data matrix $X$, $|\Omega|$ is the total number of missing values, $\hat{X}$ is the imputed complete data matrix, and $\sigma^2_{X_{\Omega}}$ is the variance of missing values. To address the bias of NRMSE with the MNAR missing mechanism, SOR has been proposed as [20]

$$\mathrm{SOR} = \sum_{i=1}^{P} rank(\mathrm{NRMSE}_i),$$

where $P$ is the number of proteins containing at least one missing value, $i$ is the protein index in this protein subset, and $rank(\mathrm{NRMSE}_i)$ is the ranks of protein-wise NRMSE across different imputation methods.

### Introduction to FRMF method

Low-rank matrix factorization is a popular and effective approach for missing data imputation [13]. For imputing proteomics data, the assumption is that there is only a small number of biological processes determining the expression profiles. This fundamental assumption is biologically plausible because measured abundances in a given sample contain substantial information of other unobserved proteins, and information of other samples with shared properties can also be useful in the learnings step [4,11]. Consider an $m \times n$ complete data matrix $\boldsymbol{X}$ describing $m$ samples and $n$ proteins. A low-rank matrix factorization approach seeks to approximate $\boldsymbol{X}$ containing missing values by a linear latent variable model,

$$\boldsymbol{X}_{m\times n} = \boldsymbol{A}_{m\times l} \times \boldsymbol{S}_{l\times n}, \#(1)$$

where $\boldsymbol{A}_{m\times l}$ and $\boldsymbol{S}_{l\times n}$ are the low-rank factor matrices, and $l \ll \min(m,n)$. To prevent overfitting, the solution is often formulated as a regularized sparse SVD minimization problem on the observed values

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} I\big(X_{ij} \neq \mathrm{NA}\big)\big(X_{ij} - \boldsymbol{A}_i\boldsymbol{S}_j\big)^2 + \lambda_{\mathrm{A}} \|\boldsymbol{A}\|_F^2 + \lambda_{\mathrm{S}} \|\boldsymbol{S}\|_F^2, \#(2)$$

where $\|.\|_F^2$ denotes the Frobenius norm, $I(.)$ is the indicator function, and $\lambda_A, \lambda_S > 0$ are the regularization parameters. When local similarity information is available, FRMF can be reformulated by adding a fused regularization term

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} I\left(X_{ij} \neq \mathrm{NA}\right)\left(X_{ij} - \boldsymbol{A}_i\boldsymbol{S}_j\right)^2 + \lambda_A \|\boldsymbol{A}\|_F^2 + \lambda_S \|\boldsymbol{S}\|_F^2 + \alpha \sum_{i=1}^{m} \sum_{k\in\boldsymbol{\mathcal{F}}(i)} \|A_i - A_k\|_F^2 , \#(3)$$

where $\alpha$ is the fused regularization parameter, and $\boldsymbol{\mathcal{F}}(i)$ denotes the neighborhood sample subset of sample $i$ and can be determined using baseline data or other relevant measurements such as gene expression or pathological score. In our study, $\boldsymbol{\mathcal{F}}(i)$ was determined by the between-sample cosine similarity $\cos(X_i, X_k)$ based on data matrix in FRMF_self, or $\cos(P_i, P_k)$ based on external pathological scores on samples in FRMF_cross_patho.

Because pair-wise local similarity among samples has already been exploited for determining neighborhood $\boldsymbol{\mathcal{F}}(i)$, we adopted the average-based fused regularization [13]. A local minimum of the objective function given by Equation 3 can be found by performing gradient descent in latent variable vectors $\boldsymbol{A}_i$ and $\boldsymbol{S}_j$,

$$\frac{\partial Err}{\partial \boldsymbol{A}_i} = \sum_{j=1}^{n} I\left(X_{ij} \neq \mathrm{NA}\right)\left(X_{ij} - \boldsymbol{A}_i\boldsymbol{S}_j\right)\boldsymbol{S}_j + \lambda_A \boldsymbol{A}_i + \alpha \sum_{k\in\boldsymbol{\mathcal{F}}(i)} \left(A_i - A_k\right), \#(4)$$

$$\frac{\partial Err}{\partial \boldsymbol{S}_j} = \sum_{i=1}^{m} I\left(X_{ij} \neq \mathrm{NA}\right)\left(X_{ij} - \boldsymbol{A}_i\boldsymbol{S}_j\right)\boldsymbol{A}_i + \lambda_S \boldsymbol{S}_j . \#(5)$$

While single-omics missing value imputation methods have been extensively studied [4,11], the research of multi-omics integrated missing value imputation strategy is relatively recent. The contributions of the work presented here are three-fold: (1) we elaborate how additional information on the sample samples can benefit missing value imputations; (2) we coin the term Fused Regularization to represent the local similarity constraints on imputation functions, and we mathematically illustrate how to design a matrix factorization objective function with fused regularization; and (3) the proposed method is quite general and can be easily extended to incorporate other contextual information such as pathological or clinical scores, etc.

### Introduction to CAM method

CAM is a latent variable modeling and deconvolution technique previously used for identifying biologically-interpretable cell subtypes or biological archetypes $\boldsymbol{S}_{l\times n}$ and their composition $\boldsymbol{A}_{m\times l}$ in complex tissue ecosystems [6,14,15,22]. We adopted the CAM framework into Equation 1 and demonstrate that hybrid CAM_SVT and CAM_NIPALS can effectively

manage missing values and that this combination leads to a novel and biologically-plausible imputation strategy. The workflow of CAM based method with three variants is given in **Figure 9**. Below we present a brief introduction to the CAM principle. Detailed mathematical descriptions and algorithms can be found in our previous publications [14,15].

The functions of complex tissues are orchestrated by a productive interplay among many specialized cell subtypes or task archetypes [32]. These biological components interact with each other to create a unique physiological or pathophysiological state. To characterize this ground yet dynamic state, mathematical deconvolution of bulk tissue data has been used to model biological tissues as an aggregate of distinct cell or molecular subtypes. The primary objective of mathematical deconvolution is to computationally detect subtype-specific markers, determine the number of constituent subtypes, calculate subtype proportions in individual samples, and estimate subtype-specific expression profiles [33]. Supported by advanced machine learning algorithms and proven theorems, unsupervised deconvolution methods can decompose the mixed molecular signals into many latent variables; these subtypes are biological interpretable and functionally enriched [6,14,34,35].

The CAM pipeline is built on the strong parallelism between nonnegative latent variable models and the theory of convex sets (**Figure 10**) [14,36,37]. Tissue samples to be modeled contain an unknown number and varying proportions of molecularly distinctive subtypes (**Figure 10a**). Molecular expression in a specific subtype is modeled as being linearly proportional to the abundance of that subtype. We showed that the scatter simplex of bulk data is a rotated and compressed version of the scatter simplex of subtype expressions (**Figure 10b**). According to the theory of convex sets [36], every molecular feature within the scatter simplex can be uniquely determined by the nonnegative combination of the vertices. Thus, the number of the vertices corresponds to the number of molecularly distinctive subtypes present in the bulk samples and the molecular features residing at the vertices are the molecular markers defining such subtypes [14]. CAM works by detecting the vertices of the scatter simplex geometrically, *i.e.*, determining the multifaceted simplex that most tightly encloses the globally measured expression mixtures. Subsequently, the molecular markers residing at the vertices are first identified, and the proportions and specific expression profiles of constituent subtypes are then estimated [14]. The number of latent components is determined by the minimum description length (MDL) criterion, given by

$$MDL(k) = \frac{1}{2}\log\left(\sum_{j=1}^{n} \|\boldsymbol{x}(j) - \mathbf{A}\mathbf{s}(j)\|_2^2\right) + \frac{(k-1)m}{2} log(n_{MG}) + \frac{kn}{2} log(m), \#(6)$$

where $k$ is the number of latent components, and $n_{MG}$ is the number of marker proteins.

## Data availability

The scripts used in the paper is available in R script ProImput. Code for all experiments can be found in the vignette at https://github.com/MinjieSh/ProImput. The operation system can be any system supporting R language.

# References


1 Canterbury, J. D., Merrihew, G. E., MacCoss, M. J., Goodlett, D. R. & Shaffer, S. A. Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics. *Journal of The American Society for Mass Spectrometry* **25**, 2048-2059 (2014).

2 Doerr, A. DIA mass spectrometry. *Nature methods* **12**, 35 (2014).

3 Goeminne, L. J. E., Sticker, A., Martens, L., Gevaert, K. & Clement, L. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Anal Chem* **92**, 6278-6287, doi:10.1021/acs.analchem.9b04375 (2020).

4 Ma, W., al., e. & Wang, P. DreamAI: algorithm for the imputation of proteomics data. *bioRxiv* (2020).

5 Dabke, K., Kreimer, S., Jones, M. R. & Parker, S. J. A Simple Optimization Workflow to Enable Precise and Accurate Imputation of Missing Values in Proteomic Data Sets. *J. Proteome Res.* **20**, 3214-3229 (2021).

6 Herrington, D. M. *et al.* Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation* **137**, 2741-2756, doi:10.1161/CIRCULATIONAHA.118.034365 (2018).

7 Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research* **15**, 1116-1125 (2016).

8 Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol* **17**, 162, doi:10.1186/s12874-017-0442-1 (2017).

9 Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief Bioinform*, doi:10.1093/bib/bbaa112 (2020).

10 Webb-Robertson, B.-J. M. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research* **14**, 1993-2001 (2015).

11 Wang, S. *et al.* NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res* **48**, e83, doi:10.1093/nar/gkaa498 (2020).

12 Lin, X. & Boutros, P. C. Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics* **21**, 7, doi:10.1186/s12859-019-3312-5 (2020).

13 Ma, H., Zhou, D., Liu, C., Lyu, M. R. & King, I. in *The fourth ACM international conference on Web search and data mining.* 287-296 (ACM Press).

14 Wang, N. *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific Reports* **6**, 18909, doi:10.1038/srep18909 (2016).

15 Chen, L. *et al.* debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. *Bioinformatics* **36**, 3927-3929, doi:10.1093/bioinformatics/btaa205 (2020).
16 Rahman, S. A., Huang, Y., Claassen, J., Heintzman, N. & Kleinberg, S. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of biomedical informatics* **58**, 198-207 (2015).
17 Pedersen, A. B. *et al.* Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology* **9**, 157 (2017).
18 John, C., Ekpenyong, E. J. & Nworu, C. C. Imputation of missing values in economic and financial time series data using five principal component analysis approaches. *CBN Journal of Applied Statistics* **10**, 51-73 (2019).
19 Cai, J.-F., Candès, E. J. & Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* **20**, 1956-1982 (2010).
20 Wei, R. *et al.* Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports* **8**, 1-10 (2018).
21 Teo, G. *et al.* mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *Journal of proteomics* **129**, 108-120 (2015).
22 Parker, S. J. *et al.* Identification of Putative Early Atherosclerosis Biomarkers by Unsupervised Deconvolution of Heterogeneous Vascular Proteomes. *J Proteome Res* **19**, 2794-2806, doi:10.1021/acs.jproteome.0c00118 (2020).
23 Lazar, C. *imputeLCMD: A collection of methods for left-censored missing data imputation*, <https://cran.r-project.org/package=imputeLCMD> (2015).
24 Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat Methods* **9**, 8-9; author reply 9, doi:10.1038/nmeth.1830 (2011).
25 Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. in *ACM SIGKDD International Conference on Knowledge Discovery & Data MiningJuly* 2623–2631 (ACM, 2019).
26 Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 611-622 (1999).
27 Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164-1167 (2007).
28 Ochoa-Muñoz, A. F., González-Rojas, V. M. & Pardo, C. E. Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm. *DYNA* **86**, 249-257 (2019).
29 Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).
30 Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112-118 (2012).
31 Oba, S. *et al.* A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19**, 2088-2096 (2003).
32 Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590, doi:10.1126/science.aaf1204 (2016).
33 Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969-1979, doi:10.1093/bioinformatics/bty019 (2018).

34 Hart, Y. *et al.* Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat Methods* **12**, 233-235, doi:10.1038/nmeth.3254 (2015).
35 Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168-1178, doi:10.1038/ng.3398 (2015).
36 Chan, T.-H., Ma, W.-K., Chi, C.-Y. & Wang, Y. A convex analysis framework for blind separation of non-negative sources. *IEEE Trans Signal Processing* **56**, 5120-5134 (2008).
37 Chen, L. *et al.* Tissue-specific compartmental analysis for dynamic contrast-enhanced MR imaging of complex tumors. *IEEE Trans Med Imaging* **30**, 2044-2058, doi:10.1109/TMI.2011.2160276 (2011).

# Figure legends

**Figure 1**. Comparative assessment of eight representative missing value imputation methods, divided into three categories.

**Figure 2**. Two-phased workflow of realistic simulation-based assessment on missing value imputation methods.

**Figure 3**. The overall pattern of missing values illustrated by the relationship between protein missing rate and protein mean intensity, before (left panel) and after (right panel) introducing masked NA.

**Figure 4**. Imputation performance of the eight methods on the simulation data of setting #1, with assumed MNAR missing mechanism and varying total missing rates.

**Figure 5**. Imputation performance of the eight methods on the simulation data of setting #1, with assumed MCAR missing mechanism and varying total missing rates.

**Figure 6**. Imputation performance of the eight methods on the simulation data of setting #2, focusing on authentic missing mechanism and varying masked rates.

**Figure 7**. Imputation performance of the FRMF variants on the simulation data of setting #2, with varying masked rates.

**Figure 8**. Imputation performance of CAM variants on the simulation data of setting #2, with varying masked rates, in comparison to that of Mean, SVT, and NIPALS. The imputation accuracy is evaluated in the original intensity space (before log-transformation).

**Figure 9**. Workflow of the CAM based imputation method with two variant algorithms.

**Figure 10**. CAM principles for latent variable modelling and deconvolution. (a) Mixed expression profile of latent process mixtures. (b) Illustration of mixing operation in scatter space, where a compressed and rotated scatter simplex whose vertices host marker genes is produced and corresponded to mixing proportions. (c) Mathematical description of expression profile of latent process mixtures.

## Acknowledgement

This work has been supported by the National Institutes of Health under Grants HL111362-05A1, HL133932, NS115658-01, and the Department of Defense under Grant W81XWH-18-1-0723 (BC171885P1).

### Author contributions

M.S. and Y.T.C. developed FRMF framework; C.T.W. and M.S. developed CAM framework; M.S. and Y.W. wrote the manuscript; M.S. and C.T.W. implemented software and performed real data analysis; S.J.P and G.S. provided datasets and technical support; D.M.H., J.E.V.E. and R.C. interpreted results and edited the manuscript; Y.Z.W. and G.Y. provided statistical expertise support. All authors have discussed the work, and read, edited, and accepted the final manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

***Supplementary Information*** The online version contains supplementary material available at …

# Figures

Figure 1.

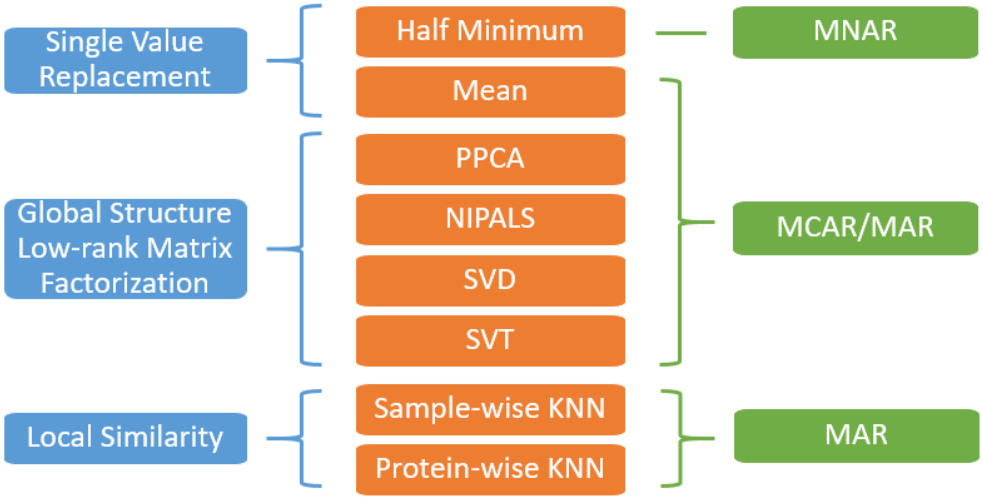
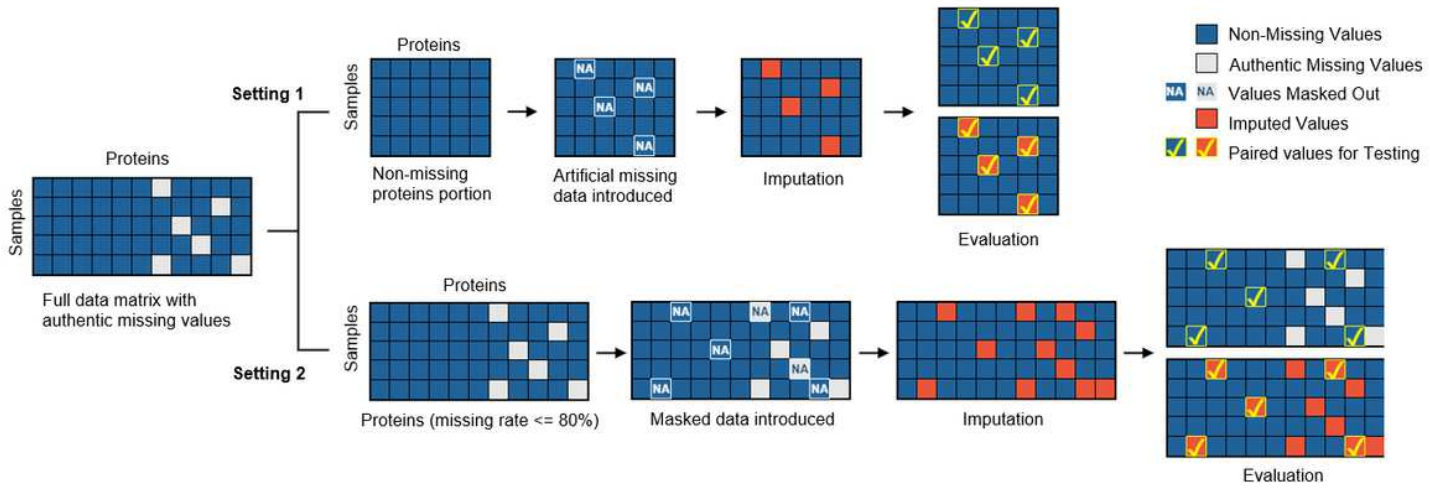



**Figure 1**

Comparative assessment of eight representative missing value imputation methods, divided into three categories.

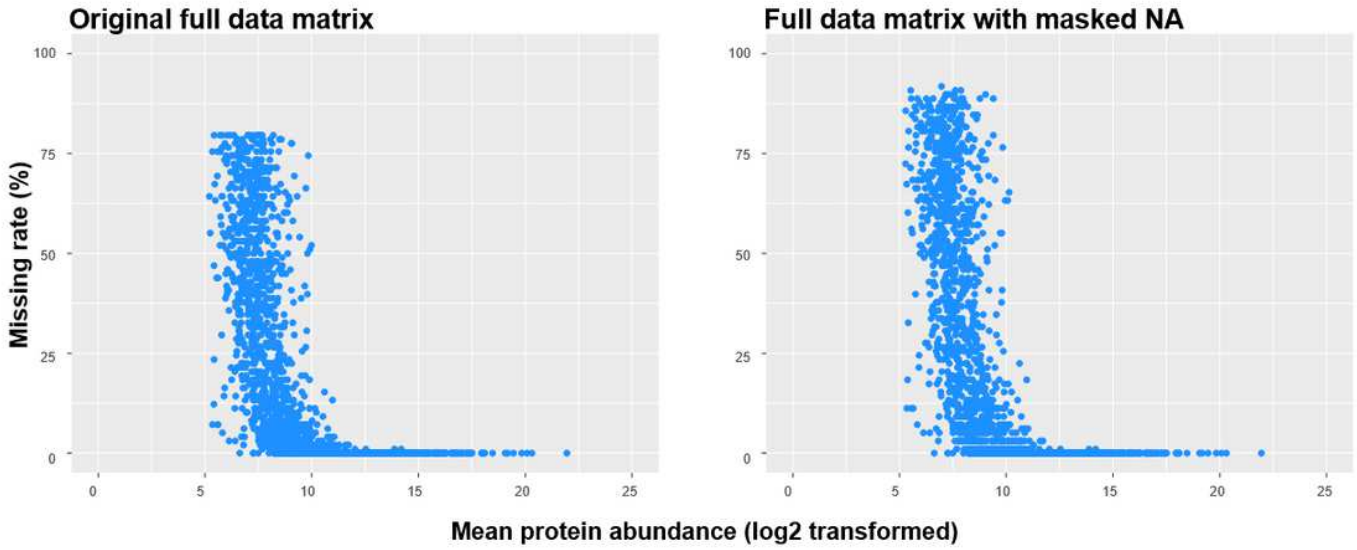
Figure 2.




## Figure 2

Two-phased workflow of realistic simulation-based assessment on missing value imputation methods.

Figure 3.




## Figure 3

The overall pattern of missing values illustrated by the relationship between protein missing rate and protein mean intensity, before (left panel) and after (right panel) introducing masked NA.

Figure 4.

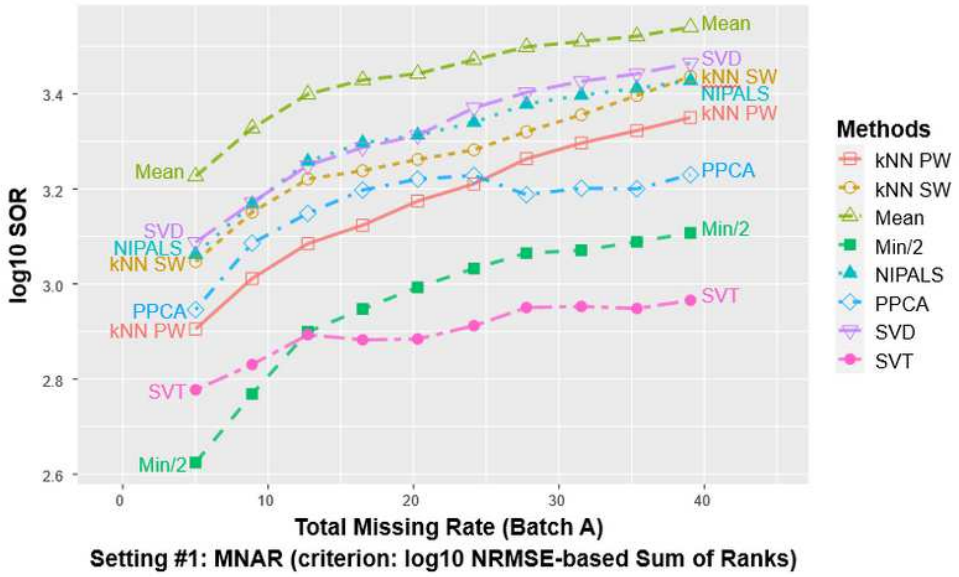



**Figure 4**

Imputation performance of the eight methods on the simulation data of setting #1, with assumed MNAR missing mechanism and varying total missing rates.

Figure 5.

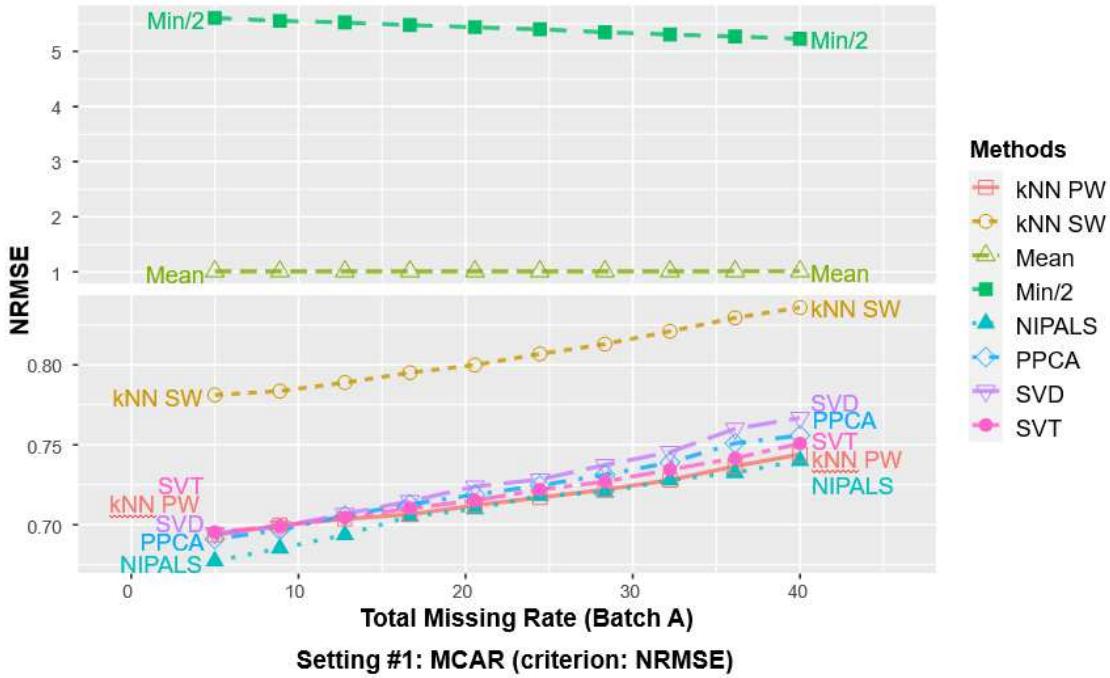



**Figure 5**

Imputation performance of the eight methods on the simulation data of setting #1, with assumed MCAR missing mechanism and varying total missing rates.

Figure 6.


## Figure 6

Imputation performance of the eight methods on the simulation data of setting #2, focusing on authentic missing mechanism and varying masked rates.

Figure 7.


**Figure 7**

Imputation performance of the FRMF variants on the simulation data of setting #2, with varying masked rates.

Figure 8.

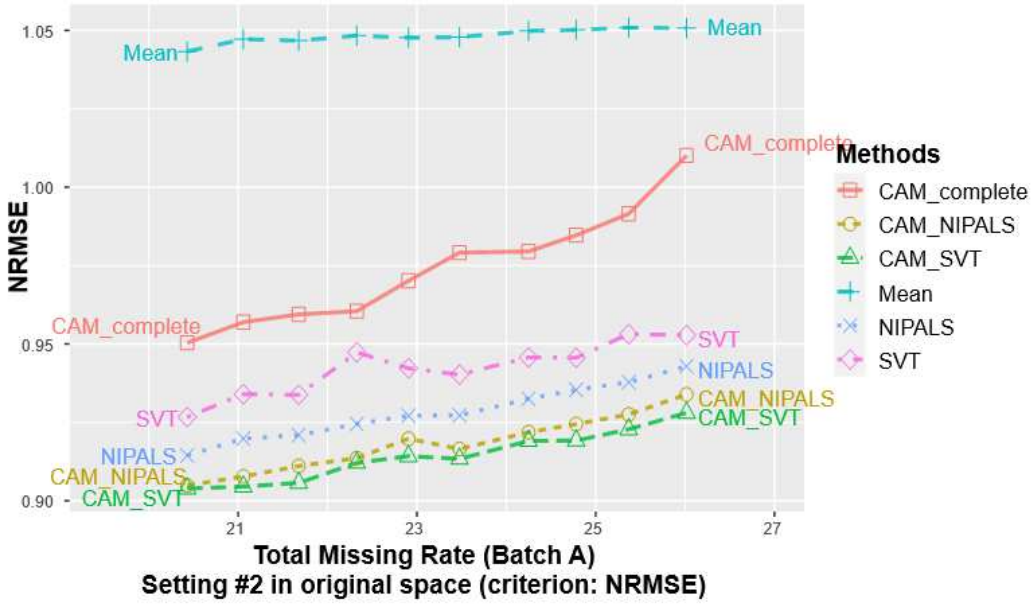



**Figure 8**

Imputation performance of CAM variants on the simulation data of setting #2, with varying masked rates, in comparison to that of Mean, SVT, and NIPALS. The imputation accuracy is evaluated in the original intensity space (before log-transformation).

Figure 9.

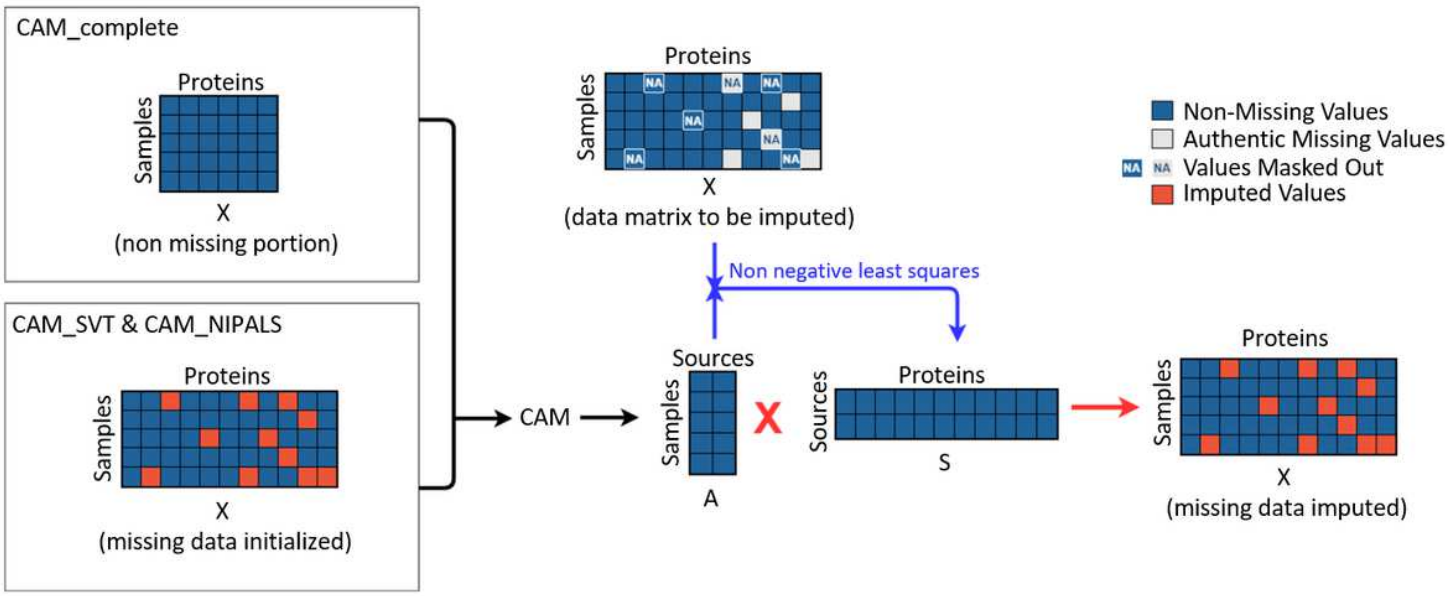



**Figure 9**

Workflow of the CAM based imputation method with two variant algorithms.

Figure 10.

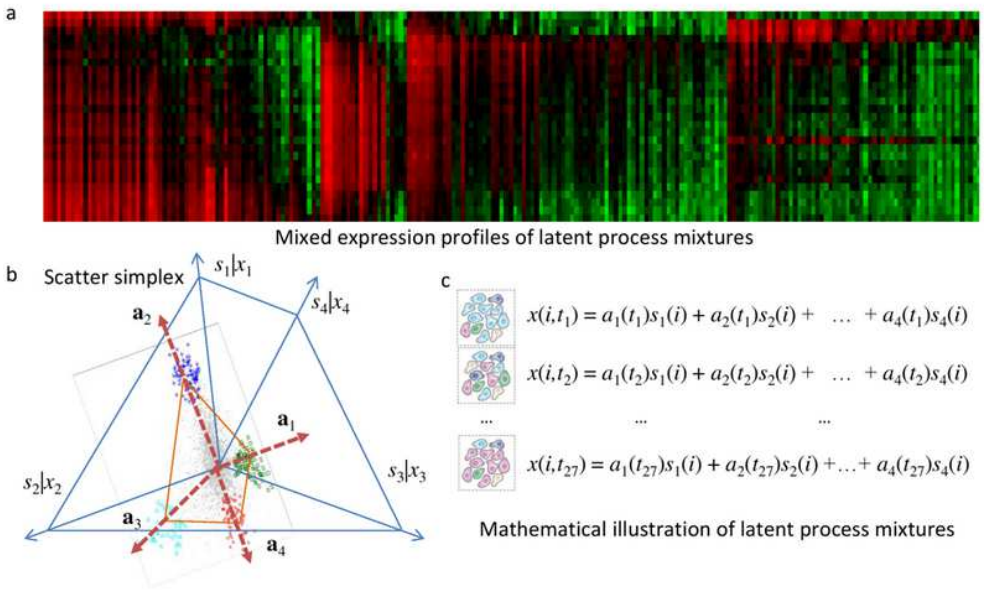



**Figure 10**

CAM principles for latent variable modelling and deconvolution. (a) Mixed expression profile of latent process mixtures. (b) Illustration of mixing operation in scatter space, where a compressed and rotated scatter simplex whose vertices host marker genes is produced and corresponded to mixing proportions. (c) Mathematical description of expression profile of latent process mixtures.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ProImputSI82021.docx