

Data and text mining

Imputation for Lipidomics and Metabolomics (ImpLiMet): a web-based application for optimization and method selection for missing data imputation

Huiting Ou^{1,2,†}, Anuradha Surendra^{3,†}, Graeme S.V. McDowell³, Emily Hashimoto-Roth^{4,5,6,7},
Jianguo Xia^{1,8}, Steffany A.L. Bennett^{4,5,9,10}, Miroslava Cuperlovic-Culf^{3,5},

¹Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada

²Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan

³Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON K1K 4P7, Canada

⁴Neurolipidomics Laboratory and India Taylor Lipidomic Research Platform, University of Ottawa, Ottawa, ON K1H 8M5, Canada

⁵Department of Biochemistry, Microbiology, and Immunology and Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada

⁶Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada

⁷Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada

⁸Institute of Parasitology, McGill University, Montreal, QC H9X 3V9, Canada

⁹Department of Cellular and Molecular Medicine, University of Ottawa Brain and Mind Research Institute, Ottawa, ON K1H 8M5, Canada

¹⁰Department of Chemistry and Biomolecular Sciences, Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding authors. Digital Technologies Research Centre, National Research Council, 1200 Montreal Rd, Ottawa, ON K1K 4P7, Canada. E-mail: miroslava.cuperlovic-culf@nrc-cnrc.gc.ca (M.C.-C.) and Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, 451 Smyth Rd, Ottawa, ON K1H 8M5, Canada. E-mail: sbennet@uottawa.ca (S.A.L.B.)

[†]Equal contribution.

Associate Editor: Guoqiang Yu

Abstract

Motivation: Missing values are prevalent in high-throughput measurements due to various experimental or analytical reasons. Imputation, the process of replacing missing values in a dataset with estimated values, plays an important role in multivariate and machine learning analyses. The three missingness patterns, including missing completely at random, missing at random, and missing not at random, describe unique dependencies between the missing and observed data. The optimal imputation method for each dataset depends on the type of data, the cause of the missingness, and the nature of relationships between the missing and observed data. The challenge is to identify the optimal imputation solution for a given dataset.

Results: ImpLiMet: is a user-friendly web-platform that enables users to impute missing data using eight different methods. For a given dataset, ImpLiMet suggests the optimal imputation solution through a grid search-based investigation of the error rate for imputation across three missingness data simulations. The effect of imputation can be visually assessed by histogram, kurtosis, and skewness, as well as principal component analysis comparing the impact of the chosen imputation method on the distribution and overall behavior of the data.

Availability and implementation: ImpLiMet is freely available at <https://complimet.ca/shiny/implimet/> and <https://github.com/complimet/ImpLiMet>.

1 Introduction

Missing data are a major problem for multivariate, machine learning (ML) and network analyses. For example, in large lipidomic or metabolic datasets, measurements for some analytes may not be available in every sample due to routine technical variability, low abundance, ion suppression from co-eluting analytes, inaccurate feature assignment in annotation pipelines, or because analytes are simply not present in a subset of samples. This “missingness” confounds ML approaches, limits the number of methodologies that can be utilized, and reduces the statistical power of models that

exclude samples with missing values. Sample exclusion further alters cohort representation, notably when “missingness” is an indicator of a particular subgroup, biasing results toward the groups in which all analytes are observed, and potentially leading to inaccurate interpretations (Jager *et al.* 2021).

Imputing missing values is commonly employed when performing multivariate and ML analyses to help reduce data bias resulting from sample exclusion. Three types of missingness have been conceptualized that can be addressed by imputation (Mack *et al.* 2018, Scheffer 2002):

- 1) Missing completely at random (MCAR) refers to values whose absence is completely independent of any other data feature or covariate. In this type of missingness, each sample has the same probability of presenting an MCAR value because there is no underlying difference between the samples with or without missing data (Rubin 1976, Mack *et al.* 2018). A real-world example of MCAR is transient (aka random) technological failure over the course of data collection such that there is no relationship between the samples with missing or observed values.
- 2) Missing at random (MAR) refers to missing values whose absence is related to the values of other measured features but not to the measured values of the same feature (Schafer 1997). Here, missing values do not depend on the variable in question but on the values of the other analytes present in each sample. An example of MAR would be when the value for one analyte is missing because its measurement is obscured by the abundance of another analyte in the same sample (e.g. ion suppression of co-eluting analytes in the case of lipidomic or metabolomic datasets).
- 3) Missing not at random (MNAR) refers to missing values that are absent because a feature, condition, or covariate is directly responsible for the absence in that sample. Here, the probability of missingness depends on the sample itself. A biological example of this group would be analytes that are not synthesized, and thus not present, in every condition. A technological example would be when analytes are present in a given sample but are below the limit of quantification of the technology used to measure the data.

Multiple imputation methods have been introduced to approximate missing values. Recently, Jäger *et al.* (2021), and Chilimoniuk *et al.* (2024), have compared and evaluated different approaches with respect to the quality of the imputed data and their downstream impact on ML pipelines. They presented a method for testing imputation quality based on the error rate and downstream use of data and in their work show that in almost all assessed examples, Random Forest (RF) provides the optimal result. To ensure agnostic dataset evaluation, Lin *et al.* (2024) have recently presented a platform for imputation of mass spectrometry omics data that provides users with the information about the hypothetical source of missingness through correlation analysis—testing possibility for MAR and statistical analysis—and exploring the possibility for missing through MNAR mechanisms. Users can then provide the ratio of missingness types present in their datasets that will influence the selection of the imputation method; however, the same imputation method is used for all variables. A remaining bioinformatic challenge is the identification of the optimal imputation solution for a given dataset of any type. As missingness can come from different sources for variables within the dataset, different imputation methods might be necessary for groups of features within the dataset.

To address this challenge, we present **Imputation for Lipidomics and Metabolomics—ImpLiMet**—applicable to any numerical dataset, validated here for using lipidomic and metabolomic data. ImpLiMet is an R package available at <https://github.com/complimet/ImpLiMet> and online through a web interface at Computational Lipidomics and

Metabolomics: CompLiMet: <https://complimet.ca/shiny/implimet/>. ImpLiMet enables users to impute missing data using eight different methods across the whole dataset or within user-defined groups of features. The effect of each method can be visualized by histogram, kurtosis, and skewness analyses, as well as principal component analysis (PCA) comparing the impact of simply removing features and samples with missing data to the chosen imputation method. To identify the optimal imputation solution, ImpLiMet further offers an optimization option wherein the error of each imputation method is evaluated, and the user is informed of the method with the lowest mean absolute percentage error (MAPE) across three “missingness” simulations for their dataset.

2 Methods

ImpLiMet is written in R and deployed as a RShiny application. Figure 1 presents the ImpLiMet workflow and pseudo-code for the optimization procedure. ImpLiMet accepts a .CSV file as input. If the dataset includes features measured in different units by different platforms (multiple feature measurement groups) or features possibly having different levels of relationships to other features, the user has the option to format their data such that the imputation methods consider feature groups separately. An example of different measurement groups could be the combination of lipidomic and metabolomic data measured using different platforms or multiomics data such as metabolomic and transcriptomic data contained in a single dataset. The user can specify the number of features or samples with the selected percentage(s) % of missing values to be removed prior to choosing an imputation measure or optimizing across measures. Eight imputation methods are available: (1) replacing with the feature minimum, (2) replacing with the feature minimum divided by 5, (3) replacing with the feature maximum, (4) replacing with the feature median, (5) replacing with the feature mean, (6) using K-Nearest Neighbors (kNN) (Hastie *et al.* 2000, Troyanskaya *et al.* 2001), (7) using RF (Pantanowitz and Marwala 2009), or (8) using Multivariate Imputation by Chained Equations (MICE) (van Buuren and Groothuis-Oudshoorn 2011). For kNN, RF, and MICE, users can specify the number of neighbors for kNN, the number of trees for RF, and the number of iterations for MICE. kNN is implemented using *impute.kNN* function; RF imputation utilizes *missRanger.RF* function (Stekhoven and Buehlman 2011) and MICE using the function *mice* (van Buuren and Groothuis-Oudshoorn 2011).

If the user’s dataset has at least 3 features and 6 samples with no missing values, or a minimum of 18 non-missing values across minimum of 3 features and 6 samples, ImpLiMet further offers an optimization option wherein the error of each imputation method is evaluated by simulating the three different sources of missingness in the user’s dataset once all missing data is removed then testing all available imputation methods. Optimization suggests the best imputation method as the one with the lowest MAPE across the three “missingness” data simulations, i.e. the lowest value for all tested values. The selected approach is used to impute the original dataset and this result is provided as a download. Alternatively, the user can choose to utilize another imputation method based on, for example, simulation results, the visualization analysis provided by ImpLiMet, or prior information about the sources of missingness in the dataset.

Figure 1. (A) Schematic workflow of ImpliMet. In the case of automated optimization, ImpliMet first removes all columns in the dataset with missing values then simulates missing elements following three types of missingness: MCAR, MAR, and MNAR. Missing values are imputed with all methods and the error of imputation is determined using MAPE. Imputation is then performed on the original dataset using the method with the lowest MAPE value. The dataset with imputed values is returned to the user and the effect of imputation on the dataset is visualized with statistical measures and PCA. (B) Schematic pseudocode of the process of data removal for the three different missingness types during optimization. Matrix multiplication indicates the element-wise product. Detailed pseudocode is provided in the [Supplementary Materials](#). A comprehensive flowchart is presented at: <https://complimet.ca/shiny/implimet/>.

In the case of different types of missingness in the dataset, the user can group features by missingness type, perform imputation using the proposed optimal methods for each group and subsequently combining the results for different groups using the downloaded data.

In the optimization step, samples without any missing values are selected to create a complete set. If the cleaned dataset obtained by removing all samples (rows) with missing values has no remaining values, optimization will instead select features (columns) without missing values. Finally, if both approaches result in the removal of all columns and rows, ImpliMet will select columns and features with <80% missing values and returns to selecting samples with no missing values with the remaining set. If not found, ImpliMet will select features with no missing values. In this way, the algorithm ensures that the analysis of the optimal imputation method for the dataset can be evaluated by imputing only the missing data from the set that is removed for testing in the optimization step. Note, if there are less than at least 18 values, in 6 samples and 3 features remaining, optimization of imputation cannot be done. It is important to keep in mind that in extremely small datasets imputation will be biased by available information. From the dataset devoid of missing features, ImpliMet removes data values at the sample threshold percentage initially provided by the user for filtering. If threshold percentage is not provided, i.e. user opts not to remove any additional features or samples from their dataset prior to imputation, ImpliMet uses 30% as the threshold percentage in the optimization process. The threshold percentage is used to simulate the optimal imputation method

for a given dataset at the level of the user's specified tolerance for imputation. For extremely small dataset sizes (e.g. a 6 × 3 matrix), only a 10% threshold for full optimization will enable simulation as all other thresholds will result in an insufficient sample size for imputation method testing and error calculation. The known values removed for simulation are kept as the hold-out set and are used to evaluate error of imputation as follows:

Given dataset: $X = [x_{ij}]_{i=1 \dots N_s, j=1 \dots N_f}$ where N_s is the number of samples and N_f is the number of features; with missing elements x_{km} ; $i=1 \dots N_s, j=1 \dots N_f$. The goal of imputation is to determine values for the missing elements that resemble the complete data. As the first step in optimization, any row or column with missing elements are removed leading to the subset $X' = [x'_{ij}]_{i=1 \dots N'_s, j=1 \dots N'_f}$ where N'_s and N'_f are the number of samples and features remaining.

From this subset data, removal is performed separately to simulate MAR, MCAR, and MNAR mechanisms. Pseudocode for each missingness mechanism is provided in [Fig. 1B](#).

For **MCAR**, a filtering matrix of dimension $N'_s \times N'_f$ is created by random sampling from a uniform distribution (minimum 0 and maximum 1) generated from the function *runif* in R. Random values in the matrix are ranked and values below the imputation threshold are set to NA for missing and above are set to one for remaining. The element-wise product between this filtering matrix and full data matrix provides the MCAR example set for further testing.

For **MNAR**, the missing value assignment is performed individually for each feature as follows: (1) A list of values is generated by sampling from a logistic distribution

(location i_{ij}^0 , scale i_{ij}^1), denoted $L_1 = \{i_{ij}^0, i_{ij}^1, \dots, N^0\}$. (2) A second list is generated by sampling from the uniform distribution (minimum i_{ij}^0 and maximum i_{ij}^1), denoted $L_2 = \{i_{ij}^2, i_{ij}^3, \dots, N^0\}$. (3) A third list is generated from the product of $L_1, L_2, L_3 = \{i_{ij}^3, i_{ij}^4, \dots, N^0\}$. (4) The ranks for the values in L_1, L_3 , as well as the feature measurements, are computed. (5) The highest and lowest ranks from L_3 , with the number of missing values dependent on the assigned threshold, are determined and the corresponding (feature-wise) ranks in L_1 are assigned. Equivalent ranks in the dataset are removed as missing.

For **MAR**, a co-dependence group is created by summing all feature values in a sample except the values in the current cell. If the input file contains information about the feature groups, based on biological or analytical characteristics, the summation calculation is performed within each feature group for each sample for the co-dependence matrix. The MAR process follows MNAR steps 1 through 3. In step 4, the ranks for the values in L_1, L_3 , and the sample values in the filtering matrix are computed. Missing indices are assigned to the highest and lowest ranks from L_3 , with the number of missing values dependent on the sample threshold. The order of the values in L_1 , which produces the missing indices in L_3 , are retrieved, and the corresponding order in the filtering matrix column for the co-dependent feature are assigned as NA.

After generating the three types of missing datasets, each dataset is imputed using each of the eight available methods. For multivariate methods, users are prompted to select a simple or full version of parameter optimization. Simple parameter optimization uses the following default parameters: K -value i_{ij}^0 10, Tree Value i_{ij}^0 500, and Mice Iteration i_{ij}^0 2. If a full parameter search is selected, the accuracy of the imputed values is tested over a range of hyperparameters for kNN, MICE and RF. Specifically, for kNN, the K -values tested range from 10 to 100 incremented by 20. For the optimal K -value in this range, a refined search is conducted from $k - 4$ to $k + 4$ in single value increments to identify the K -value with the lowest error rate. For RF, the number of trees in the sequence of 5, 10, 20, 50, 100, 150, 200, 500 are examined to determine the optimal tree size. For MICE, 1–3 iterations are tested. The full optimization approach is generally preferred, however due to the large number of calculations taken in this approach it can be time consuming (e.g. for dataset with 45 samples 40 features—the example input set provided—full optimization test takes 2 min online). Thus, for very large datasets, fast optimization analysis can provide initial screen of methodologies. Error rates are calculated by mean absolute error rate (MAPE) defined as:

$$MAPE = \frac{100}{N} \sum_{i=1}^N \frac{|x_i - y_i|}{x_i}; \quad x_i > 0; \quad (1)$$

where N is the number of missing values, x_i is the actual value, and y_i is the prediction. The MAPE results for each of the eight imputation methods assessed for each missingness mechanism are displayed and the method with the lowest MAPE value across the missingness mechanisms is highlighted and used for imputation. In general, omics measurements are greater than zero as the minimal value measured

corresponds to the minimal level of detection in the measurement, rather than absolute zero value.

The effect of imputation on the dataset is visualized by dataset histogram, kurtosis, and skewness characteristics as well as PCA comparing the original dataset, following removal of all samples and features with missing data, to that of the imputed dataset. Histograms show all values in the dataset following feature z-score scaling and compares the overall dataset distribution of cleaned dataset with the imputed set. Kurtosis and skewness provide information about the distribution for each feature separately. Kurtosis is a measure the level of tailing of the data. Skewness indicates the symmetry relative to the normal distribution. Symmetric data has a skewness of zero. High negative skewness indicates that data are left skewed (a long-left tail, thus data are missing more values in the high abundance range). Positive skewness indicates data are right-skewed, meaning that more low abundance data are missing altering the assumption of a normal distribution. High skewness, calculated in ImpLiMet using R function *skewness*, suggests the possibility of MNAR for those features. Kurtosis (calculated using R function *kurtosis*), indicates potential increased levels of outliers in the dataset, with high values suggesting significant presence of outliers from normal distribution. In ImpLiMet, kurtosis and skewness are shown for both datasets with all samples and features with missing values removed and the complete, imputed dataset, allowing the user to explore possibility for of MNAR in some of the features as well as to observe the effect of imputation on the normality of features distribution. PCA, for both samples, calculates principal components using features as variables, and displays features, using their values across samples as variables. The user-provided sample and feature names are shown in the plots for reference. An example of the optimization utilization as well as comparison of errors in imputation using recommended and other imputation methods is presented in the [Supplementary Materials](#).

Briefly, from the subset of metabolomics data published by [Li et al. \(2019\)](#) with complete data for 50 samples and 50 features, we have removed values from 120 cells and tested the error rate for the imputed values using different methods. Results show that the recommended method, in this case RF, provides imputation with the lowest error and the best agreement in PCA when comparing the original dataset with the original data (full information is provided in [Supplementary Materials](#)). We also provide an example of the utilization of ImpLiMet on a combined metabolomics and lipoprotein dataset ([Oppong et al. 2024](#)) with multiple groups and testing of the skewness analysis ([Supplementary Materials](#)).

3 Results

ImpLiMet is a versatile and web-based application designed to assist users in identifying the optimal imputation solution for their datasets. It identifies the optimal method based on the lowest error rate overall, while at the same time presenting error rates of imputation for different types of missingness for all methods. ImpLiMet currently includes eight imputation methods as well as visual representation of statistical features of the dataset to help users interpret sources of missingness across features. Future development will include the addition of other imputation methods as well as an automated analysis of the type of missingness present in the data.

Author contributions

Huiting Ou (Methodology [equal], Software [lead], Writing—original draft [equal], Writing—review & editing [equal]), Anuradha Surendra (Formal analysis [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [lead], Writing—review & editing [equal]), Emily Hashimoto-Roth (Software [equal], Visualization [equal], Writing—review & editing [equal]), Graeme S.V. McDowell (Methodology [equal], Software [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jianguo Xia (Resources [equal], Supervision [equal], Writing—review & editing [equal]), Steffany A.L. Bennett (Conceptualization [equal], Funding acquisition [equal], Investigation [equal], Project administration [equal], Resources [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Miroslava Cuperlovic-Culf (Conceptualization [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

[Supplementary data](#) are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported in part by RGPIN-2019-06796 to S.A.L.B. from the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as operating grant AI-4D-102-3 to S.A.L.B. and M.C.-C. from the National Research Council of Canada AI4Design Program. H.O. received an NSERC CREATE Matrix Metabolomics Scholarship.

Data availability

There are no new data associated with this article.

References

- Chilimoniuk J, Grzesiak K, Kała J *et al.* Imputomics: web server and R package for missing values imputation in metabolomics data. *Bioinformatics* 2024;**40**:2024.
- Hastie T, Tibshirani R, Eisen MB *et al.* ‘Gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* 2000;**1**:research0003.
- Jäger S, Allhorn A, Bießmann F. A benchmark for data imputation methods. *Front Big Data* 2021;**4**:693674.
- Li H, Ning S, Ghandi M *et al.* The landscape of cancer cell line metabolism. *Nat Med* 2019;**25**:850–60.
- Lin W, Ji J, Su KJ *et al.* omicsMIC: a comprehensive benchmarking platform for robust comparison of imputation methods in mass spectrometry-based omics data. *NAR Genom Bioinform* 2024;**6**:lqae071.
- Mack C, Su Z, Westreich D. *AHRQ Methods for Effective Health Care, Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User’s Guide*, 3rd edn. Rockville (MD): Agency for Healthcare Research and Quality (US), 2018.
- Oppong AE, Coelewijn L, Robertson G *et al.* Blood metabolomic and transcriptomic signatures stratify patient subgroups in multiple sclerosis according to disease severity. *iScience* 2024;**27**:109225.
- Pantanowitz A, Marwala T. Missing data imputation through the use of the random forest algorithm. In: Yu W, Sanchez EN (eds), *Advances in Computational Intelligence. Advances in Intelligent and Soft Computing*, Vol. **116**. Berlin, Heidelberg: Springer, 2009.
- Rubin DB. Inference and missing data. *Biometrika* 1976;**63**:581–92.
- Schafer JL. *Analysis of Incomplete Multivariate Data*, 1st edn. Boca Raton, USA: Chapman and Hall/CRC, 1997.
- Scheffer JL. Dealing with missing data. *Res Lett Inf Math Sci* 2002;**3**:7.
- Stekhoven DJ, Bühlmann P. MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2011;**28**:112–8.
- Troyanskaya O, Cantor M, Sherlock G *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**:520–5.
- van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Soft* 2011;**45**:1–67.