Bioinformatics Advances, 2025, **00**, vbae209
https://doi.org/10.1093/bioadv/vbae209
Advance Access Publication Date: 21 January 2025
Application Note

OXFORD

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. McDowell[3], Emily Hashimoto-Roth[4,5,6,7], Jianguo Xia [1,8], Steffany A.L. Bennett [4,5,9,10,*], Miroslava Čuperlović-Culf [3,5,*]

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada
[2]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan
[3]Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON K1K 4P7, Canada
[4]Neurolipidomics Laboratory and India Taylor Lipidomic Research Platform, University of Ottawa, Ottawa, ON K1H 8M5, Canada
[5]Department of Biochemistry, Microbiology, and Immunology and Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada
[6]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada
[7]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada
[8]Institute of Parasitology, McGill University, Montreal, QC H9X 3V9, Canada
[9]Department of Cellular and Molecular Medicine, University of Ottawa Brain and Mind Research Institute, Ottawa, ON K1H 8M5, Canada
[10]Department 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 (Jäger *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):

Received: August 14, 2024; **Revised:** November 26, 2024; **Editorial Decision:** December 22, 2024; **Accepted:** January 17, 2025

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.

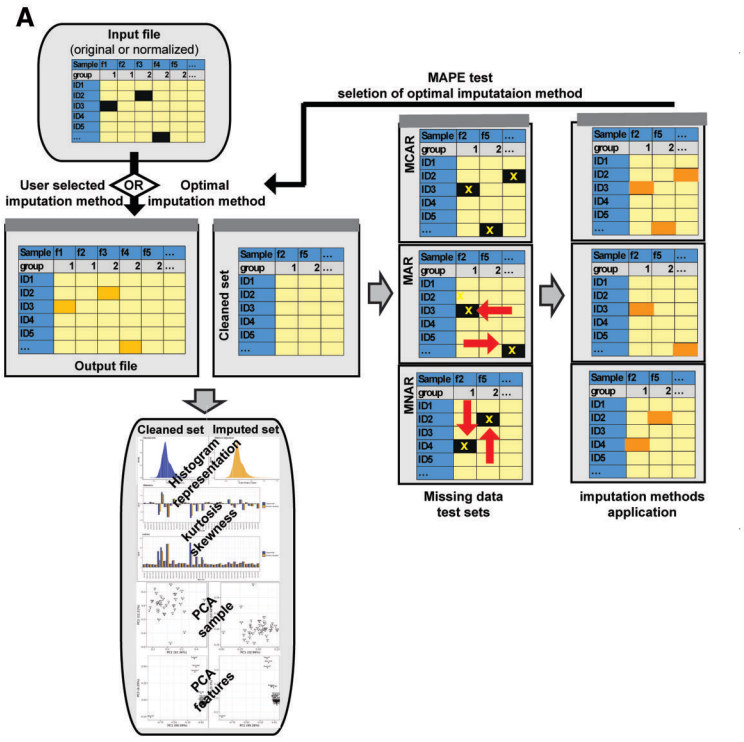
Multiple imputation methods have been introduced to approximate missing values. Recently, Jäger *et al.* (2021), and Chilimoniuk *et al.* (2024), have compared and evaluated different approaches with respect to the quality of the imputed data and their downstream impact on ML pipelines. They presented 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 possibly 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 **Imp**utation for **Li**pidomics and **Met**abolomics—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 pseudocode 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.



B

MCAR dataset := set with removed values

Filtering.matrix:= matrix of random numbers
ranked.values=rank values filtering.matrix
If ranked.value < threshold
filter.matrix =NA
data subset = dataset · filter matrix
return dataset

MAR dataset := set with removed values

Filtering.matrix:= matrix of values dependent on row values (measurements of other features)
ranked.values=rank values filtering.matrix
If ranked.value <> threshold
filter.matrix =NA
data subset = dataset · filter matrix
return dataset

MNAR dataset := set with removed values

Filtering.matrix:= matrix of values dependent on the column values (measurements for the feature in other samples)
ranked.values=rank values filtering.matrix
If ranked.value <> threshold
filter.matrix =NA
data subset = dataset · filter matrix
return 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: $X = \{x_{ij}\},\ i = 1\ldots,\ N_s; j = 1\ldots,\ N_f$ where N_s is the number of samples and N_f is the number of features; with missing elements $x_{km},\ (k,m) \in M$ 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 $X' = \{x_{ij'}\},\ i = 1\ldots,\ N_s',\ x_{km},\ (k,m); j = 1\ldots,\ N_f$.

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 = 0, scale = 1), denoted $L_1 = \left\{l_i^{(1)}\right\},\ i = 1\ldots,\ N's$. (2) A second list is generated by sampling from the uniform distribution (minimum = 0 and maximum = 1), denoted $L_2 = \left\{l_i^{(2)}\right\},\ i = 1\ldots,\ N's$. (3) A third list is generated from the product of $L_3 = L_1 \cdot L_2$, $L_3 = \left\{l_i^{(3)} = l_i^{(1)} l_i^{(2)}\right\}$. (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-dependance group is created by summating 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 = 10, Tree Value = 500, and Mice Iteration = 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:

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

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 Čuperlović-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, Coelewij 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.

Bioinformatics Advances, 2025, 00, 1–5
https://doi.org/10.1093/bioadv/vbae209
Application Note

Supplementary Materials

ImpLiMet: Online optimization and method selection for missing data imputation

Huiting Ou[1-2+], Anuradha Surendra[3+], Graeme S.V. McDowell[3], Emily Hashimoto-Roth[4,6-8], Jianguo Xia[1,5], Steffany A.L. Bennett[4,6,9*], Miroslava Čuperlović-Culf[3,6*]

[1]Department of Human Genetics, McGill University, Montreal, Quebec, Canada
[2]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[3]National Research Council of Canada, Digital Technologies Research Centre, Ottawa, Ontario, Canada
[4]Neurolipidomcs Laboratory and India Taylor Lipidomic Research Platform, University of Ottawa, Ottawa, Ontario, Canada
[5]Institute of Parasitology, McGill University, Montreal, Quebec, Canada
[6]Department of Biochemistry, Microbiology, and Immunology and Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Ontario, Canada
[7]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada
[8]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[9]Department of Cellular and Molecular Medicine, University of Ottawa Brain and Mind Research Institute, & Department of Chemistry and Biomolecular Sciences, Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada
+ Equal first authors *To whom correspondence should be addressed.

Contact: steffanyann.bennett@uottawa.ca and miroslava.cuperlovic-culf@nrc-cnrc.gca.

Contents

S1. Optimization method performance testing

Imputation methods included in ImpLiMet have all been previously developed, tested, and extensively used (Chilimoniuk, et al. 2024; Hastie et al. 2000; Pantanowitz and Marwala, 2009; Stekhoven et al. 2011; Troyanskaya et al. 2001; van Buuren et al. 1999; van Buuren et al. 2006; van Buuren et al. 2011; Wright and Ziegler, 2017). If *Optimization* is selected in the analysis, ImpLiMet determines the imputation error rate for different methods and suggests to the user the best performing imputation method for the dataset. The optimal method for imputation for a given dataset is performed through a grid search across all methods and with range of hyperparameters. The error level is determined for three different types of missingness: Missing completely at random (MCAR), Missing not at random (MNAR), and Missing at random (MAR). Hyperparameter values used in the optimization search are shown in Supplementary Table 1.

Supplementary Table 1. Hyperparameter values included in the optimization of machine learning imputation methods.

Method	KNN	RF	MICE
Type of Parameter	K-value	Tree value	Number of iterations
Optimization: Full search	10:100 (20) Fine search: Min-4:Min+4 (1)	5, 10, 20, 50:200 (50), 500	1:3 (1)
Optimization: Fast search	10	500	2

The range of values is shown. The step value for the list is indicated in brackets.

Imputation error is calculated by mean absolute error rate (MAPE) defined as:

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{x_i}$$

where N is the number of missing values, x_i is the actual value and y_i is the prediction.

Imputation optimization is done using a subset of data that does not have any missing values. In this case ImpLiMet first removes samples (rows) with any missing values. If the resulting subset has less than 6 rows (samples), ImpLiMet instead removes all features (columns) with any missing values from the original dataset. If the remaining set has less than 3 features and less than 6 samples, the optimization step cannot be performed. In this case, the user can still select their preferred imputation method and perform imputation on the original set. If 6 or more samples and 3 or more features in the provided set are complete, i.e., they have no missing values, then these are selected and used for the imputation method error testing and optimization. To determine the error rate, missing data are simulated by removing cells from the cleaned dataset (the previously selected samples and features without missing data) following the procedures described in Supplementary Box 1.

Supplementary Box 1. Overview, pseudo code showing methods for cell removal that represent different missingness types. All matrix products correspond to element-wise matrix products.

MCAR dataset := set with removed values

Filtering.matrix:= matrix of random numbers
ranked.values=rank values filtering.matrix
If ranked.value < threshold
filter.matrix =NA
data subset = dataset · filter matrix
return dataset

MNAR dataset := set with removed values

For i=1:number of columns
L1:= logistic distribution random set
L2:=uniform distribution random set
L3 = L1 · L2

ranked.L3 (1:row,i)=rank values L3
ranked.L1(1:row,i)=rank values L1
ranked.values(1:row,i)=
rank column values

MAR dataset := set with removed values

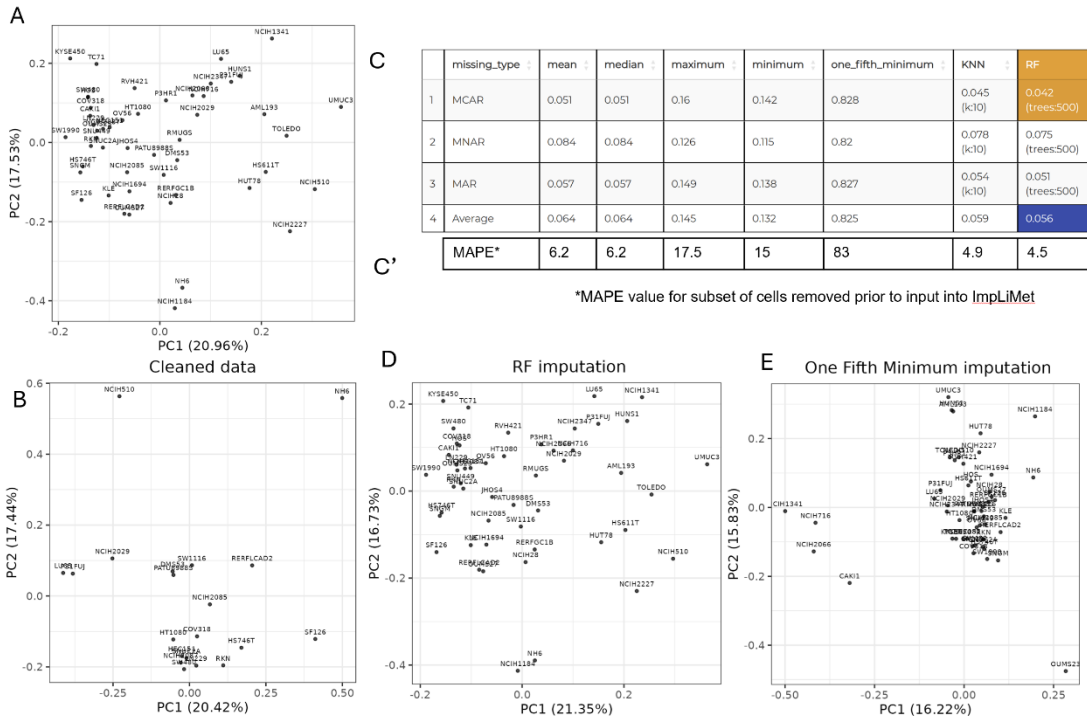
For i=1:number of columns
L1:= logistic distribution random set
L2:=uniform distribution random set
L3 = L1 · L2
$sum.row = \sum_{\forall\, row\ \setminus\{current\}} values$

ranked.L3 (1:row,i)=rank values L3
ranked.L1(1:row,i)=rank values L1

	missing.rank= *ranked.L1≅ranked.L3<>threshold* *filter.column=NA, when* *ranked.values eq missing.rank* *data subset = dataset · filter matrix* *EndFor* *return dataset*	*ranked.values(1:row,i)=* *rank sum.row* *missing.rank=* *ranked.L1≅ranked.L3<>threshold* *filter.column=NA, when* *ranked.values eq missing.rank* *data subset = dataset · filter matrix* *EndFor* *return dataset*

Missing data is imputed using all methods included in ImpLiMet. The difference between the imputed and the original values in the cleaned set is calculated using the MAPE formula. The minimal MAPE value is suggested as the optimal method and is used to impute the dataset's existing missing values. A table of MAPE values for the three different missingness approaches and all imputation methods is provided. The imputation recommendation depends on the characteristics of samples as well as type of missingness and the sample size. The optimization method is a simple grid search identifying the method that provides the lowest error rate across all missingness patterns in their specific dataset.

An example of the performance of the imputation method optimization is shown using metabolomics dataset published by Li et al. (Li et al. 2019). The subset used in this example measures 50 samples and 50 features. The selected set does not have any missing values. PCA for the complete dataset is shown in Supplementary Figure 1A. From this dataset we have removed values from 120 cells, PCA for the cleaned dataset, the subset of values with no missing value selected following random deletion of 120 values is shown in Supplementary Figure 1B. On this dataset we ran optimization and imputation with the recommended method as well as all other methods. Imputation results for different methods are compared using MAPE calculation for the imputed and original values, prior to removal, as well as comparison results using PCA results. Results are shown in the Supplementary Figure 1.



C

	missing_type	mean	median	maximum	minimum	one_fifth_minimum	KNN	RF	MICE
1	MCAR	0.051	0.051	0.16	0.142	0.828	0.045 (k:10)	0.042 (trees:500)	0.095 (iteration index:2)
2	MNAR	0.084	0.084	0.126	0.115	0.82	0.078 (k:10)	0.075 (trees:500)	0.093 (iteration index:2)
3	MAR	0.057	0.057	0.149	0.138	0.827	0.054 (k:10)	0.051 (trees:500)	0.099 (iteration index:2)
4	Average	0.064	0.064	0.145	0.132	0.825	0.059	0.056	0.096

C'

MAPE*	6.2	6.2	17.5	15	83	4.9	4.5	10.0

*MAPE value for subset of cells removed prior to input into ImpLiMet

Supplementary Figure 1. Example of optimization and imputation presented on a subset of the metabolomics data of (Li et al. 2019). Subset includes 50 samples and 50 features with no missing values. A. PCA for the complete, original set used in the analysis. From this set we randomly removed 120 values across different samples and features for imputation analysis. B. PCA of the subset with only complete rows and columns after the removal of 120 values from the set shown in A. C. For this dataset ImpLiMet optimization recommends Random Forest (RF) as the optimal imputation method as this approach leads to the lowest MAPE value overall in MCAR missingness type. C' As in this case we have the original values for the 120 cells removed in the analysis, we are also including here MAPE values for the imputed and original values, prior to removal. D. PCA of the optimized imputation result using RF, the lowest MAPE value method. E. PCA of samples following the imputation using 1/5th of the minimum value which is in this case has the highest MAPE values.

This benchmark analysis shows that error in imputation values for the missing data fully agrees with the MAPE values obtained in the optimization process. The recommended method, in this case Random Forest (RF) has the lowest error rate based on MAPE analysis using imputed values and values in the original dataset. Furthermore, PCA of the RF imputed dataset is in the agreement with the original dataset's PCA – with highly comparable PC1 and PC2 values (Supplementary Figure 1. A and D) while PCA values following 1/5th of the minimum value imputation, as a suboptimal method in this dataset, show different PC1 and PC2 values and PCA projection structure (Supplementary Figure 1. E).

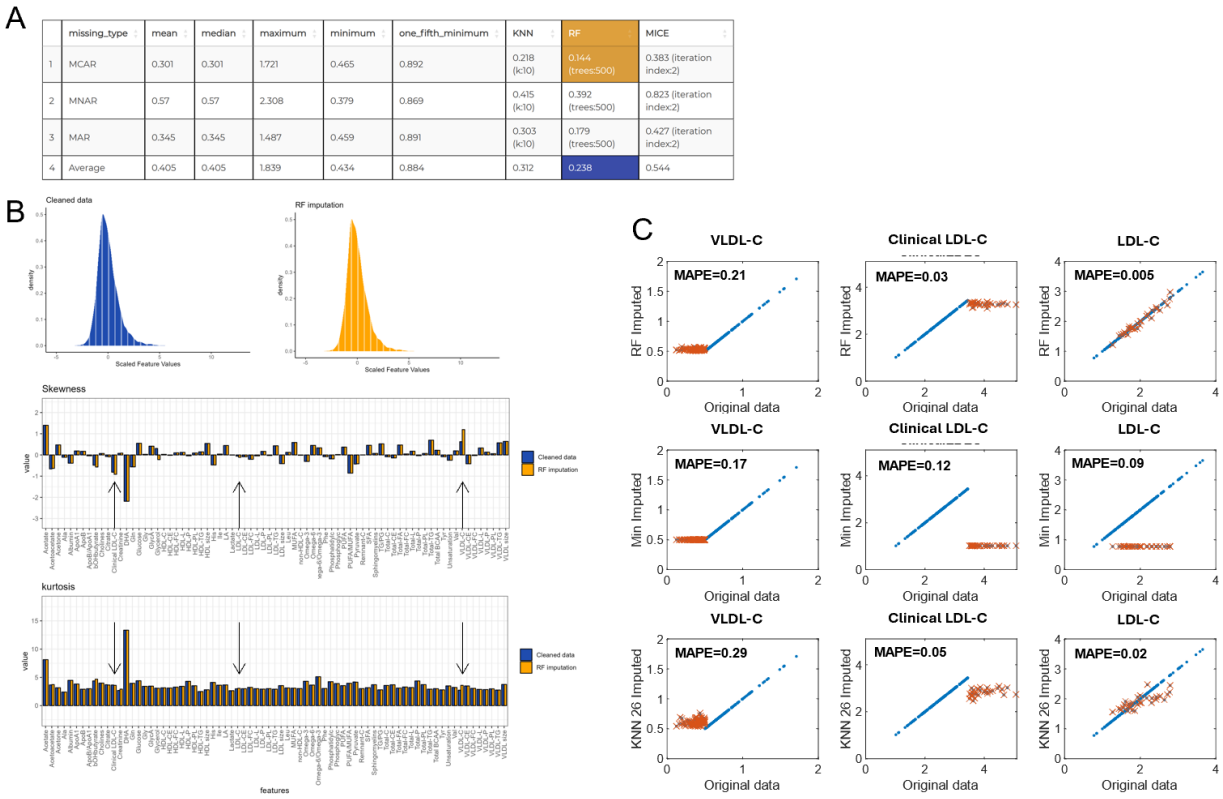
S2. Example: Metabolomics dataset imputation

In this example we show utilization of Implimet for the imputation of a dataset previously published by Oppong et al. (Oppong, 2024). Briefly, this dataset contains metabolomics and lipoprotein measurements in serum of 191 patients

with relapsing-remitting Multiple Sclerosis (MS) (RRMS, N=52), patients with neuromyelitis optica (DCs, N=30), secondary progressive MS (SPMS, N=29), and control healthy donors (HC, N=80). Included in the analysis are both metabolites and lipoproteins measured in serum using a high throughput NMR spectroscopy platform. Study design and detailed analysis are provided in the original publication. This dataset had several values set to zero, and thus, prior to imputation testing all zero values in the dataset are set to missing values. Additionally, for this application presentation here for three features we have respectively removed data that are below a threshold value, above a threshold or randomly selected through the feature set. Specific features are shown in Supplementary Figure 2. A set of 75 metabolites and lipoproteins was selected for this demonstration with metabolites and lipoproteins in this dataset separated into two groups such that imputation was performed only within a group of features (adding as a second row named "groups" with group labels). Although in this set both groups of features are measured using NMR methodology, they required two different pulse sequences for analysis of small molecules (metabolites) and large constructs (lipoproteins). Following the upload ImpLiMet shows dataset to have 191 samples, 75 features and 296 missing values. In this analysis we are selecting to not remove any samples and features prior to imputation. Optimization investigation in this dataset indicates that imputation using RF with 5 trees has the lowest MAPE value overall obtained for the MCAR missingness type. Optimization table is shown in Supplementary Figure 2A. We have selected full optimization which is using hyperparameters listed in Table 1.

A

	missing_type	mean	median	maximum	minimum	one_fifth_minimum	KNN	RF	MICE
1	MCAR	0.301	0.301	1.721	0.465	0.892	0.218 (k:10)	0.144 (trees:500)	0.383 (iteration index:2)
2	MNAR	0.57	0.57	2.308	0.379	0.869	0.415 (k:10)	0.392 (trees:500)	0.823 (iteration index:2)
3	MAR	0.345	0.345	1.487	0.459	0.891	0.303 (k:10)	0.179 (trees:500)	0.427 (iteration index:2)
4	Average	0.405	0.405	1.839	0.434	0.884	0.312	0.238	0.544



Supplementary Figure 2. Imputation analysis using ImpLiMet presented on the Multiple Sclerosis dataset published by (Oppong, 2024). A. A full parameter search analysis of the optimal method for imputation on this dataset. Indicated and used for subsequent imputation is the method that has the lowest MAPE value in all types of missingness, in this case RF with 5 trees. B. Visual outputs of some statistical characteristics of the input dataset with features with missing values removed (Cleaned data) and dataset following RF imputation. Shown are data values histograms, skewness and kurtosis analysis for each feature separately. C. For method presentation in this example we have deliberately removed values below 0.5 in VLDL-C measurements, above 3.5 in Clinical LDL-C and randomly 30 values of 191 for LDL-C. Plots compare original values for these three features with values obtained in imputation (red) as well as

all the other values (blue, unaffected by imputation). Shown are results for RF Imputation, min value imputation and KNN with 26 neighbours. Each plot shows MAPE values for the feature.

For larger dataset or faster screening, it is possible to do optimization on a single hyperparameter (without selecting the full parameter search in the options for optimization) but whenever possible it is recommended to do the full test. Supplementary Figure 2B shows some of the result visualization provided by ImpLiMet. Histogram for the cleaned data where features with missing values are removed compared to the set with RF Imputed data shows no change in the overall value distribution. As there are only 296 values missing out of total of 14325 values, this is expected. Skewness and kurtosis values are shown for each feature separately for the two datasets (Supplementary Figure 2B). For VLDL-C and Clinical LDL-C, where values are removed prior to imputation for values below or above a threshold, there is a slight increase in absolute value of skewness following imputation. Comparison of original values and values obtained with different types of imputation (Supplementary Figure 2C) clearly shows reasons for this increase in the distribution skewness where imputation leads to values around the threshold. The RF imputation shows for these specific features minimal MAPE values, with particularly low error rate in the example of LDL-C, where values have been removed completely at random. For the VLDL-C, values below concentration of 0.5 are removed, RF imputation result matches minimal value in the remaining set. For Clinical LDL-C, where values over 3.5 have been removed, RF values are largely matching the maximum remaining value, clearly leading to an increased in skewness. Thus, for VLDL-C, imputation with minimum value leads to a highly comparable MAPE value with RF imputation. Thus, although optimization analysis provided by ImpLiMet does not test for the missingness source in the dataset, analysis of skewness and kurtosis can be used by user to determine possible sources of missingness through the investigation of the left or right-side skewness of the original and imputed dataset.

S3. ImpLiMet web application

(a) Input File format

Table 2. Example of the ImpLiMet input file format required if the dataset has only one feature measurement group

Sample	feature1	feature2	feature3	feature4	...
ID1	15669.4	205.2	56.5	361.5	12
ID2	10084.3	220.9		438	8.3
ID3	12836.6	394.7	93.9	861.2	
...	10520.2	293	200.1	1309.9	

Row 1 must contain feature names. Column 1 must contain unique sample IDs. Missing values should be indicated as NA or as empty cells.

Table 3. Example of the ImpLiMet input file format required if the user includes information about multiple feature measurement groups

Sample	feature1	feature2	feature3	feature4	...
group	1	1	1	2	2
ID1	15669.4	205.2	56.5	361.5	12
ID2	10084.3	220.9		438	8.3

ID3	12836.6	394.7	93.9	861.2	
...	10520.2	293	200.1	1309.9	

If the dataset includes features measured in different units by different platforms (multiple feature measurement groups), data should be formatted to indicate which groups must be considered separately for missing data simulation (i.e., which data were measured in the same units on the same platform). In this case Row 1 must contain feature names. Row 2 must contain the group information. Column 1 must contain the sample IDs. Missing values should be indicated as NA or as empty cells.

(b) Running ImpLiMet

The user uploads the dataset, selects the percentage threshold for imputation. After threshold is selected, the user can select type of imputation method that will be used. If the optimization option is selected, user can further select the full parameter search. With this option the calculation of MAPE value for three types of missingness is performed using hyperparameters listed in Table 1.

(c) Output

The ImpLiMet output includes the imputed dataset as well as a visualization of the effect of the chosen (or optimized) imputation method on the dataset. The visualization tabs provide histograms for cleaned and imputed datasets as well as comparison of kurtosis and skewness values for each feature in the original and imputed datasets. For visualization of overall effect of imputation ImpLiMet also shows PCA plots of the cleaned dataset (i.e., dataset with all columns and rows with any missing value removed) and of dataset imputed with the selected method. PCA is performed on both samples and features with the names of samples and features included in the plot for easy reference. The optimization result is shown in the MAPE table. The method with the lowest MAPE for the dataset across three missingness types is highlighted and used for imputation.

References:

Chilimoniuk, J., et al. (2024), 'imputomics: web server and R package for missing values imputation in metabolomics data', Bioinformatics, 40 (3).

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol.1(2): research0003.

Li H, Ning S, Ghandi M, Kryukov GV, Gopal S, Deik A, Souza A, Pierce K, Keskula P, Hernandez D, Ann J, Shkoza D, Apfel V, Zou Y, Vazquez F, Barretina J, Pagliarini RA, Galli GG, Root DE, Hahn WC, Tsherniak A, Giannakis M, Schreiber SL, Clish CB, Garraway LA, Sellers WR. (2019) The landscape of cancer cell line metabolism. Nat Med.;25(5):850-860.

Oppong AE, Coelewij L, Robertson G, Martin-Gutierrez L, Waddington KE, Dönnes P, Nytrova P, Farrell R, Pineda-Torra I, Jury EC. (2024) Blood metabolomic and transcriptomic signatures stratify patient subgroups in multiple sclerosis according to disease severity. iScience.; 27(3):109225.

Pantanowitz, A., Marwala, T. (2009). Missing Data Imputation Through the Use of the Random Forest Algorithm. In: Yu, W., Sanchez, E.N. (eds) Advances in Computational Intelligence. Advances in Intelligent and Soft Computing, vol 116. Springer, Berlin, Heidelberg.

Stekhoven, Daniel J., and Peter Buehlmann. (2011) MissForest-non-parametric missing value imputation for mixed-type data." Bioinformatics 28 (1): 112–18.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, H., Tibshirani, R., Botstein, D., Altman, R.B. (2001) Missing value estimation methods for DNA microarrays , Bioinformatics, 17(6), 520–525

van Buuren, S., Boshuizen, H.C., Knook, D.L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, 681–694.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn C.G.M., Rubin, D.B. (2006) Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76, 12, 1049–1064.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67.

Wright, M., Ziegler. A. (2017) "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." Journal of Statistical Software, 77 (1): 1–17.