msBayesImpute as a versatile framework for addressing missing values in biomedical mass spectrometry proteomics data

Info

PUBLISHED DATE
2026-07-07

JOURNAL
Communications Chemistry

LINK
https://doi.org/10.1038/s42004-026-02106-3

Authors

Jiaojiao He Barbara Helm Franziska Gödtel Katharina Büchner Marcel Schilling Marc A. Schneider Laura V. Klotz Jana Braunger Hauke Winter Britta Velten **Ursula Klingmüller** Junyan Lu

Abstract

Advancements in mass spectrometry (MS) technologies have significantly improved the ability to quantify proteins and analyse their modifications. However, MS-based proteomics datasets frequently encounter missing values due to a complex interplay of missing at random (MAR) and missing not at random (MNAR) mechanisms. Such missing data can result in information loss and biased outcomes in data pre-processing, as well as subsequent analyses and interpretations. Few approaches effectively address both MAR and MNAR, and those that do often necessitate manual tuning of mixture percentages between them or rely on two-group experimental designs. Therefore, we developed msBayesImpute, an innovative computational method that integrates Bayesian factorization with probabilistic dropout models. We evaluated msBayesImpute against several popular imputation methods using both simulated missing values and those generated through a dilution series experiment on samples from lung cancer patients. Our comprehensive benchmark demonstrated superior performance in reconstructing missing values, estimating normalization factors, identifying differentially expressed proteins and predicting outcomes with machine learning models across varying levels of missingness and sample sizes. Notably, msBayesImpute does not require predefined experimental designs and is scalable to large-scale studies. This versatility positions msBayesImpute as an effective and robust tool for enhancing the utility of MS datasets in biological research.