

PRISM: A Proteomics Robust Imputation framework for Structure-aware Modeling of missingness

Tiannan Guo

guotiannan@westlake.edu.cn

Westlake University <https://orcid.org/0000-0003-3869-7651>

Zhaoxing Li

Westlake University

Yi Chen

Westlake University

Zhiwen Yang

Yangzhou University <https://orcid.org/0009-0009-1608-7554>

Article

Keywords: proteomics, mass spectrometry, deep learning, imputation, missing values, autoencoder, matrix factorization

Posted Date: September 4th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7480159/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Yes there is potential Competing Interest. T.G. is a shareholder of Westlake Omics Inc.

PRISM: A Proteomics Robust Imputation framework for Structure-aware Modeling of missingness

Zhaoxing Li^{1,2,3*}, Zhiwen Yang^{1,2,3,4†}, Yi Chen^{1,2,3*}, Tiannan Guo^{1,2,3*}

¹Affiliated Hangzhou First People's Hospital, State Key Laboratory of Medical Proteomics, School of Medicine, School of Future Biomedicine, Westlake University, Hangzhou, Zhejiang Province, China;

²Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang Province, China;

³Research Center for Industries of the Future, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China.;

⁴School of Mathematics, Yangzhou University, Yangzhou, China.

[†]Work performed while intern at Guomics.

*Corresponding author: lizhaoxing@westlake.edu.cn, chenyi@westlake.edu.cn, guotianan@westlake.edu.cn

Abstract

1 Missing values in mass spectrometry (MS)-based proteomics, particularly within label-free quantification
2 (LFQ) workflows, hinders downstream data analysis. These missing values are predominantly categorized
3 as Missing Not At Random (MNAR), as they often arise when the signals of low-abundance proteins or
4 peptides fall below the instrument's limit of detection. To address this, we introduce PRISM, a novel
5 imputation framework comprising two complementary deep learning models: a Denoising Convolutional
6 Autoencoder (DCAE) and a Deep Matrix Factorization (DMF). The core innovation of the framework is
7 its explicit modeling of the intrinsic MNAR missingness mechanism through a gradient downsampling
8 strategy and a dual-task learning objective. We conducted a comprehensive evaluation on multiple
9 real-world and synthetic proteomics datasets, assessing performance on metrics including numerical
10 accuracy (RMSE), downstream classification task performance, and preservation of clustering structure.
11 The results demonstrate that PRISM not only surpasses existing methods in imputation accuracy but, more
12 importantly, also excels at preserving the inherent biological structure of the data, thereby supporting more
13 accurate downstream prediction and exploratory analysis. Therefore, PRISM provides the proteomics
14 research community with a more accurate toolkit that better preserves the original structure of data, thereby
15 enhancing the credibility of biological discoveries obtained from incomplete data.

16 **Keywords:** proteomics; mass spectrometry; deep learning; imputation; missing values; autoencoder;
17 matrix factorization.

1 Introduction

Mass spectrometry (MS)-based proteomics has become an indispensable tool in modern biological research, enabling large-scale identification and quantification of proteins from complex biological samples to drive significant advancements in fields such as biomarker discovery, systems biology, and personalized medicine¹. However, a persistent challenge in this technology, particularly in label-free quantification (LFQ) workflows, is the prevalence of missing values in the resulting protein abundance matrices^{2,3}. These missing values can significantly hinder downstream statistical analyses, leading to reduced reproducibility, decreased statistical power for detecting differentially expressed proteins, and potentially biased biological interpretations^{4,5}. Missing data mechanisms are broadly categorized as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR)⁶. A substantial body of evidence suggests that missing values in LFQ proteomics are predominantly MNAR, often arising because proteins or peptides of low abundance fall below the instrument's limit of detection (LOD) or limit of quantification (LOQ)^{7,8}. This intensity-dependent missingness is particularly problematic, as simply ignoring these missing values or applying naive imputation strategies can lead to severe bias in the data structure, mask true biological signals, or introduce artificial differences between sample groups^{2,9}. Therefore, addressing MNAR appropriately is crucial for accurate quantitative proteomics data imputation.

To address the pervasive issue of missing values, a diverse array of imputation strategies has been proposed¹⁰⁻¹². Foundational approaches can be broadly grouped into three main categories^{13,14}: single-value replacement, which substitutes all missing entries with a single constant or with random draws from a simple univariate distribution; methods based on local similarity, which impute missing values by leveraging information from a subset of similar data points using distance metrics or local regression models such as KNN^{15,16}; and methods based on global similarity, which utilize information from the entire dataset to model overall data structure by identifying underlying patterns or latent variables^{17,18}.

Recently, deep learning methods, especially autoencoders, have revolutionized the analysis of biological data due to their powerful capacity for non-linear modeling, achieving significant successes within proteomic field¹⁹⁻²¹. Given their potential, deep learning has also been explored for the general tasks such as data imputation^{22,23}. Various architectures, such as Variational Autoencoders (VAEs), Importance-Weighted Autoencoders (IWAEs)²⁴ and Collaborative Filtering (CF)^{25,26} have proven effective for imputation. For instance, MIWAE²⁷ is primarily applied to handle MAR scenarios, while not-MIWAE²⁸ offers a novel approach for handling MNAR data by explicitly modeling the missingness process. Building on the success of deep learning in general imputation tasks, several models have rapidly emerged that are specifically designed for proteomics data²⁹⁻³². These methods represent the current state-of-the-art proteomic data imputation methods.

However, despite the multitude of existing imputation methods, several challenges persist within the field. Firstly, most proteome imputation methods, even advanced deep learning models, do not explicitly model the inherent Missing Not At Random (MNAR) mechanism, which is driven by instrument detection limits, as a source of information. Secondly, current proteome imputation methods focus predominantly on the numerical accuracy of recovering artificially created missing values, overlooking a more critical issue: imputed data may not necessarily preserve the complex structural features of the original dataset. This neglect of structural fidelity may incur downstream biological analyses, such as prognostic modeling or clustering, often yield highly heterogeneous and non-robust results, casting doubt on the practical utility of imputation itself³³. What's more, while missing value imputation is a common practice in the single-cell transcriptomics and broader machine learning communities, algorithms specifically tailored for proteomics data are comparatively scarce. Therefore, the field urgently needs a new paradigm that not only pursues numerical precision but is also committed to preserving biological structure, validated by its performance in complex downstream tasks.

To address these challenges, we introduce PRISM, a novel framework designed to provide a more accurate, biologically relevant, and generalizable imputation solution for proteomics data. The PRISM framework comprises two models that tackle the imputation task from different perspectives (see Figure 1, Methods). The first is a model based on a denoising convolutional autoencoder (DCAE), which treats each sample as an independent vector and uses a convolutional network to learn complex, non-linear relationships among protein features to reconstruct the original signal from a corrupted input. The second model is based on deep matrix factorization (DMF), which learns a low-dimensional latent embedding for each sample and each protein individually. It then utilizes a multilayer perceptron (MLP) to combine a given sample and protein embedding to predict the corresponding expression value. By offering these two models with different inductive biases, PRISM provides users with a flexible and

70 comprehensive imputation toolkit.

71 2 Results

72 2.1 Framework Overview and Dataset Preparation

73 The overall workflow of this study is illustrated in Figure 1. Our PRISM framework integrates two deep learning
 74 models and employs a dual-task learning strategy (see Methods). During training, both the DMF and DCAE
 75 models are required not only to reconstruct the original expression values but also to predict the missingness mask
 76 generated by our gradient downsampling strategy. This design compels the models to understand the intrinsic MNAR
 77 missingness patterns of the data, which acts as a powerful regularizer. This ensures the resulting imputations are not
 78 only more numerically accurate but also have higher fidelity in preserving biological structure.

79 Our proposed DMF model enhances the standard matrix factorization (MF) approach with two primary innovations.
 80 Instead of the linear dot product used in traditional MF, we first employ a Multi-Layer Perceptron (MLP) to learn
 81 the non-linear interactions between sample and protein embeddings for expression prediction. The core novelty,
 82 however, is the introduction of a second, parallel MLP tasked with modeling the missing data mechanism itself. By
 83 combining the reconstruction loss and a masking loss into a joint objective function, our model learns embeddings
 84 that are simultaneously optimized for both data reconstruction and an understanding of the underlying missingness
 85 patterns.

86 The advantage of our DCAE model, on the other hand, lies in its use of a convolutional neural network as the encoder.
 87 Compared to a fully-connected MLP, a convolutional architecture has lower parameter complexity and computational
 88 load, and its kernel-based design can effectively capture local correlations between features within a proteomic
 89 expression profile.

90 To comprehensively evaluate the performance and generalizability of the PRISM framework, we first curated a
 91 diverse suite of real-world clinical proteomics datasets. Our selection criteria ensured that these datasets varied
 92 sufficiently in terms of sample origin, sample size, number of features, and original missingness rate. The detailed
 93 statistics of these datasets are summarized in Table 1. However, evaluation on real-world data alone does not provide
 94 a ground truth for the missing values, making it difficult to precisely quantify numerical imputation accuracy. To
 95 address this, we additionally generated a high-fidelity synthetic dataset. This dataset, modeled after a real breast
 96 cancer cell line dataset using a Copula-based method (see methods), was created to serve as a gold standard, a dataset
 97 that is statistically consistent with real data but is itself entirely complete—thereby providing a controllable and
 98 reliable benchmark for subsequent performance evaluations.

99 As shown in Supplementary Figure S1, we randomly selected several proteins and visualized their distributions in
 100 both the original and synthetic datasets. The distribution curve of the synthetic data highly overlaps with that of the
 101 original data. This provides strong visual evidence that our generation method successfully replicated the marginal
 102 distribution of each individual protein, including its original skewness and kurtosis characteristics.

103 For quantitative confirmation, we further performed two-sample t-tests and Kolmogorov-Smirnov (KS) tests on all
 104 features between the original and synthetic datasets. None of the tests showed a significant difference ($p > 0.05$).
 105 Together, this visual and statistical evidence confirms that our generated synthetic data is a "digital twin" that is
 106 statistically consistent with the real data, allowing it to serve as a reliable gold standard for the subsequent evaluation
 107 of imputation algorithms.

108 2.2 Analysis of imputation accuracy and model robustness

109 To thoroughly evaluate the performance of the two models within our proposed PRISM framework, we tested their
 110 accuracy and efficiency under a series of controlled conditions.

111 First, we investigated the effect of the missingness prediction task as a regularizer on model performance. In our
 112 dual-task learning framework, the total loss consists of a data reconstruction loss and a missingness prediction
 113 loss, with the latter being modulated by a weight λ . As shown in Table 2 and Table 3, we calculated the RMSE
 114 on artificially mask data points for DMF and DCAE with λ varies from 0.1 to 1.0. The results showed that the
 115 models' RMSE reached its lowest point when λ is set to 0.4. This indicates that forcing the model to simultaneously
 116 reconstruct the data and learn the mask generated by our gradient downsampling strategy—which simulates realistic

MNAR patterns effectively improves the performance of the primary task of imputation. This task acts as an regularizer, guiding the model to learn the intrinsic regularities of missingness, rather than simply fitting the observed data points. This enhances the model's generalization ability and the final quality of the imputation.

Next, we conducted a benchmark of the two models against various methods across multiple datasets. As shown in Table 4, with a fixed λ weight, the results clearly demonstrate that the RMSE of our DCAE and DMF models is significantly lower than that of existing methods on the vast majority of datasets. This proves that the PRISM framework surpasses many current mainstream and advanced algorithms in terms of imputation accuracy.

2.3 Evaluating the impact of imputation on downstream classification

To further validate the practical utility of different imputation methods in addressing real biological questions, we conducted a series of downstream supervised learning tasks. We assessed the performance of machine learning models trained on data imputed by each method to predict key clinical endpoints. Specifically, we performed three independent binary classification tasks on three different datasets: 1) predicting Biochemical Recurrence (BCR) on the PCSHA dataset; 2) distinguishing between benign and malignant tumors on the TPD dataset; and 3) predicting lymphatic metastasis on the ZZOV dataset. For each task, the imputed feature matrix was first reduced to 50 dimensions using principal component analysis (PCA), and a simple Random Forest classifier was subsequently trained on the data imputed by different methods. We used classification accuracy as the primary evaluation metric.

As shown in Table 5, the results indicate that data imputed by the PRISM framework supports the construction of superior classification models in most cases. On PCSHA and ZZOV datasets, PRISM achieved the highest classification accuracy, outperforming other benchmark methods. However, it is particularly noteworthy that in the analysis of the TPD dataset, every imputed dataset performed worse in classifier construction than the original matrix with simple zero-filling. This finding implies that while advanced imputation methods can be effective, they do not guarantee improved performance in every downstream analysis, and the intrinsic characteristics of the dataset itself play a crucial role. Nevertheless, the overall superior performance of PRISM across a majority of diverse downstream tasks provides strong evidence of its remarkable ability to recover and preserve the critical biological signals necessary for clinical predictive modeling.

2.4 Preservation of Data Structure in Downstream Clustering Analysis

A superior imputation algorithm should not only perform well in supervised learning tasks but should also maximally preserve the original intrinsic structure of the data, which is crucial for unsupervised exploratory analyses. To evaluate this, we assessed the clustering performance on the PTV1 dataset after imputation with the different methods.

As shown in Figure 2, the evaluation results clearly point to one conclusion that algorithms based on matrix factorization are superior at preserving the original clustering structure of the data. Our PRISM-DMF and the NMF model achieved the highest scores on key clustering-related metrics, particularly in the Adjusted Rand Index (ARI) and k-Nearest Neighbors classification accuracy (kNN-accuracy), as they directly measure the concordance between the clusters formed from the imputed data and their true biological labels.

This finding suggests that matrix factorization-based algorithms, whether linear like NMF or our deep non-linear DMF, may possess an inductive bias that is particularly well-suited for separating biological signals. By decomposing high-dimensional data into a combination of separate parts, they can effectively capture and preserve the core biological signatures that define the identity of different cell lines. Therefore, when the research goal involves tasks that rely on the clear clustering of samples, an imputation strategy based on matrix factorization may be the optimal choice.

3 Discussion

In this study, we proposed and comprehensively validated a novel dual-model imputation framework PRISM to address the pervasive issue of missing values in proteomics data. Benchmark tests on multiple datasets demonstrate that our models generally outperform various existing mainstream methods in terms of imputation accuracy. More importantly, our study goes beyond simple numerical comparisons to demonstrate the superiority of PRISM through a series of downstream analyses. In downstream clinical endpoint prediction tasks, data imputed by PRISM supports

the construction of higher-performing classification models. Concurrently, in exploratory clustering analysis of the data's intrinsic structure, our methods, particularly the matrix factorization-based DMF model, show the strongest ability to preserve the structure of biological groups. Taken together, these results collectively indicate that PRISM not only provides more numerically accurate imputations but also better preserves the data structure crucial for downstream biological discovery, offering a more reliable and comprehensive solution to this long-standing challenge in the field.

The core of PRISM's success lies in its specifically designed structure to handle missing values in proteomics. It ceases to treat imputation as a mere mathematical problem of filling in the blanks, and instead leverages the missingness event itself as a valuable information to be learned and utilized. Traditional imputation methods, even those based on an MNAR assumption, merely apply a static, global rule. This one-size-fits-all strategy inevitably distorts the original variance and covariance structure of the data, thereby corrupting the biological patterns that are crucial for downstream analysis.

In contrast, PRISM was trained through gradient downsampling strategy, exposes the model to a dynamic MNAR environment during training that simulates the real-world physical process. The model is forced to learn the intrinsic rule that lower signal intensity corresponds to a higher probability of missingness. Furthermore, our dual-task learning architecture compels the model to do more than just guess a number; it must understand the context of why that number is missing. Optimized by a combined loss function, the model must learn to distinguish between a true low expression and a value missing due to being below the limit of detection, and to generate different, more plausible imputed values for these scenarios. It is this explicit modeling of the missingness mechanism that makes PRISM's imputations more biologically plausible, and explains why it best preserves the data structure to achieve optimal performance in downstream classification and clustering tasks.

An interesting finding of our study is that the two models within our framework, DCAE and DMF, demonstrate distinct advantages in different downstream tasks. The DCAE model, based on a convolutional autoencoder, excels at capturing local dependencies and complex non-linear patterns among protein features within an individual sample, owing to the nature of its convolutional kernels. In contrast, the DMF model, based on deep matrix factorization, follows a collaborative filtering logic by learning global latent embeddings for each sample and each protein to model their interactions. This also indicates that there is no universal imputation method that can be applied to all proteomics datasets. It is recommend to select different methods based on the actual data structure before use.

It is particularly noteworthy that in the clustering task, which aims to preserve intrinsic biological groupings, the matrix factorization-based methods performed exceptionally well. We speculate this is because matrix factorization is fundamentally designed to learn a parts-based representation. In datasets like PTV1, where the dominant source of variation is the distinct identity of different cell lines, the unique proteomic profile of each cell line can be considered an independent part. Matrix factorization algorithms are highly effective at deconstructing the original data matrix into a combination of these constituent parts, thus giving them a natural advantage in tasks that require the clear separation of these biological groups. This result also reveals that there is no single one-size-fits-all model architecture that is optimal for all tasks; the choice of the best method is intimately linked to the intrinsic structure of the data and the goal of the downstream analysis.

Finally, although this study demonstrates the effectiveness of imputing at the protein level, we acknowledge that this is not the only solution. A core future direction of our work will be to extend the PRISM framework to imputation at the peptide level. We envision developing a version of PRISM capable of handling the highly sparse peptide matrix and performing co-imputation at both the peptide and protein levels. By comparing the results of protein inference after peptide-level imputation against the results of direct protein-level imputation, we expect the two approaches to be mutually consistent and cross-validated. We believe that this multi-level consistency check will provide more robust and convincing imputation results than analysis at a single level, thereby establishing a more solid data foundation for downstream biological discoveries.

4 Conclusion

Missing Not At Random (MNAR) values, driven by instrument detection limits, represent a core challenge hindering downstream analysis in proteomics research. This study introduces PRISM, a novel deep learning framework designed to address this problem by offering complementary models, including a convolutional autoencoder and a deep matrix factorization model. The core innovation of our framework is the explicit modeling of the data

missingness mechanism itself through a unique dual-task learning architecture. Comprehensive evaluations on multiple real-world and synthetic datasets demonstrate that PRISM not only surpasses various state-of-the-art methods in imputation accuracy but, more importantly, also excels at preserving the data's inherent biological structure for downstream classification and clustering tasks. Notably, our work also highlights the particular strength of matrix factorization-based approaches in preserving the signatures of distinct biological groups. We therefore believe that PRISM provides the proteomics community with a more reliable and flexible toolkit, enhancing the accuracy and credibility of downstream biological discoveries.

5 Materials and Methods

5.1 Materials

This study utilized five proteomic datasets to train and evaluate the performance of the PRISM framework.

5.1.1 HeLa Cell Line Proteome Dataset

This dataset was generated from quality control (QC) and maintenance (MNT) runs of mass spectrometers³⁴. From an initial collection of 50,521 raw files, 11,062 Data-Dependent Acquisition (DDA) runs were processed. After filtering for runs with at least 15,000 identified peptides and a minimum raw file size of 0.618 GB, 7,444 unique runs were retained for analysis. The UNIPROT human reference proteome database (2019_05 release) was used for searches.

5.1.2 Alzheimer's Disease Brain Proteome Dataset

This dataset, aimed at understanding the molecular signatures of Alzheimer's Disease (AD), comprises peptide-centric quantitative proteomics data from human brain samples³⁵. Samples were collected from four brain regions (superior and middle temporal gyri, hippocampus, inferior parietal lobule, and caudate nucleus) from individuals across a spectrum of AD pathology. The raw files were converted to mzML format, and chromatogram libraries were generated using EncyclopeDIA against a Prosit-predicted spectral library, with final peak integration and report generation performed in Skyline.

5.1.3 PTV1 Proteome Dataset

This is a large-scale perturbation proteomics dataset generated by treating 18 breast cancer cell lines with 63 FDA-approved drugs and 914 drug combinations³⁶. After quality control, 16,311 Data-Independent Acquisition (DIA) MS data files were included. Data was analyzed with DIA-NN software using a breast cancer-specific spectral library. Prior to model input, missing values were imputed with 0.8 times the minimum detected intensity across the dataset, and min-max normalization was performed on each sample, resulting in expression levels for 5,585 proteins per condition.

5.1.4 PCa Proteome Dataset

This study performed proteomic profiling on 918 Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimens from 306 Chinese patients with primary prostate cancer (PCa)³⁷. Samples for each patient included punches from tumor areas and an adjacent non-cancerous region. The DIA data from the samples were parsed using DIANN software. In preprocessing, proteins missing in over 80% of samples in both tumor and benign groups were removed, and remaining missing values were imputed using the impSeq method from the NAGuideR¹⁸ package.

5.1.5 TPD Proteome Dataset

Generated for the development of an AI-based classifier for thyroid nodules, this dataset included FFPE tissue samples and Fine-Needle Aspiration (FNA) biopsies³⁸. Raw DIA files were analyzed using DIA-NN software against a thyroid-specific spectral library. For classifier development, missing values for the 6,689 protein features were imputed with the minimum of all available feature values in the discovery set.

5.1.6 ZZOV Proteome Dataset

This dataset was generated to create a multi-histotype proteomic landscape of Epithelial Ovarian Cancer (EOC) for the discovery of diagnostic biomarkers and therapeutic targets³⁹. The study conducted a comprehensive proteomic profiling of ovarian tissue and plasma samples from 813 patients with different histotypes and therapeutic regimens, covering the expression of 10,715 proteins. For tissue samples, a Pressure Cycling Technology (PCT)-assisted PulseDIA method was used, quantifying a total of 10,527 proteins. For plasma samples, tandem mass tag (TMT)-based proteomics was employed, quantifying 1,660 proteins. During preprocessing, the tissue protein matrix underwent quantile normalization and its missing values were imputed as 0.8 times the minimum value, while the plasma protein matrix was batch-corrected using Combat.

5.1.7 Synthetic Proteome Dataset

To establish a controllable ground truth for the precise evaluation of imputation algorithms, an additional synthetic proteomics dataset was generated. This synthetic data was modeled after a breast cancer cell line DIA dataset, which was acquired in Data-Independent Acquisition (DIA) mode on an Orbitrap mass spectrometer and consisted of 94 samples and 8,040 identified proteins. Our generation process began by identifying and isolating the largest possible sub-matrix with no missing values from this real dataset.

Subsequently, a Copula-based method was employed to model the statistical properties of this complete sub-matrix^{40,41}. Specifically, for each protein feature in the sub-matrix, the optimal marginal distribution was first selected by fitting function families. A Gaussian Copula was then used to learn and capture the complex joint dependency structure among all protein features. As the correlation matrix estimated from high-dimensional data may not be strictly positive-definite—a requirement for the Cholesky decomposition used in the subsequent generation step—we implemented a correction measure: the correlation matrix was subjected to an eigen-decomposition, all negative eigenvalues were replaced with their absolute values, and a new, guaranteed positive-definite correlation matrix was then reconstructed⁴². Finally, by sampling from this corrected Copula model and applying the inverse cumulative distribution function for each feature, a new synthetic data matrix is generated that is statistically consistent with the real data but is itself entirely complete, providing a reliable gold standard for subsequent evaluation experiments. The resulting synthetic matrix showed no significant differences from the original data when compared using two-sample t-tests and Kolmogorov-Smirnov (KS) tests⁴³, confirming its high statistical fidelity.

5.2 Methods

5.2.1 Problem Definition and Preprocessing

We first need to define the problem. Assume we have a dataset with M samples and N proteins.

- $X_{\text{orig}} \in \mathbb{R}^{M \times N}$: The true, complete protein abundance matrix with no missing values.
- $M \in \{0, 1\}^{M \times N}$: The observation mask matrix, where $m_{ij} = 1$ if the original value $x_{\text{orig},ij}$ is observed, and $m_{ij} = 0$ if it is missing.
- $A \in \mathbb{R}^{M \times N}$: The observed protein abundance matrix containing missing values. If $m_{ij} = 1$, then $A_{ij} = X_{\text{orig},ij}$; if $m_{ij} = 0$, then A_{ij} is a missing value (NA).

Before feeding the data into our model, we preprocess the observed matrix A in the following two steps to obtain the final model input matrix, A_{proc} :

1. **Initial Filling:** Missing values (NA) in the observed matrix A are first replaced with 0 to obtain a filled matrix, A_{fill} .

$$A_{\text{fill},ij} = \begin{cases} A_{ij} & \text{if } o_{ij} = 1 \\ 0 & \text{if } o_{ij} = 0 \end{cases}$$

2. **Column-wise Min-Max Normalization:** Subsequently, each column j of A_{fill} is normalized using the min-max method. To prevent information leakage from the test set, the minimum and maximum values used for normalization are calculated exclusively from the **training set**. A small epsilon, ϵ , is added for numerical stability.

$$A_{\text{proj}} = \frac{A_{\text{fill},j} - \min(A_{\text{fill},j,\text{train}}) + \epsilon}{\max(A_{\text{fill},j,\text{train}}) - \min(A_{\text{fill},j,\text{train}})}$$

We denote a single preprocessed sample vector from A_{proc} as $x \in \mathbb{R}^N$.

5.2.2 Model Architecture

The PRISM framework comprises two core model architectures with different inductive biases: a model based on a denoising convolutional autoencoder (DCAE) and a model based on deep matrix factorization (DMF). Both models employ a dual-task learning strategy, simultaneously reconstructing the original data signal while also learning the data's missingness patterns.

Denoising Convolutional Autoencoder (DCAE) Model This model processes each sample as an independent vector. Its main component is a denoising convolutional autoencoder (DCAE), which is tasked with reconstructing the complete, original sample $\hat{x}$ from a partially corrupted input sample x' generated via the gradient downsampling strategy. Concurrently, a parallel missingness prediction network takes the reconstructed sample $\hat{x}$ and predicts which positions were artificially masked, outputting mask logits $\hat{m}$.

Deep Matrix Factorization (DMF) Model This model follows the logic of collaborative filtering and matrix factorization. Instead of processing the sample as a whole, it learns a separate, low-dimensional latent embedding vector for each sample and each protein in the dataset. To predict an arbitrary value x_{ij} , the model retrieves the embedding for sample i and protein j , concatenates them, and feeds the resulting vector into two parallel MLPs. One MLP is responsible for reconstructing the original protein expression value $\hat{x}_{ij}$, while the other is responsible for predicting the missingness state $\hat{m}_{ij}$ at that position.

5.2.3 Training Strategy

To train and evaluate our models under controlled conditions, we employ a unified strategy to partition training and validation sets from any given original dataset, regardless of whether it already contains missing values. Our partitioning method gradient downsampling is a probabilistic sampling strategy designed to simulate the real-world MNAR process.

Specifically, we first apply Gradient Downsampling to all observed values in the original data matrix. The data points that are sampled and removed by this process, along with their original values, constitute our validation set. These values are held out from model training and serve as the ground truth for evaluating the final imputation accuracy.

The remainder of the dataset, which now contains both the native missing values and the newly created validation missing values, serves as the basis for our training set. During the model training phase, we dynamically and repeatedly apply the gradient downsampling strategy to the still-observed values within this training set to generate the corrupted samples x' and their corresponding original samples x for model learning. The core idea of gradient downsampling is that the lower a feature's value, the lower its success probability p of being kept in a Bernoulli trial. For each feature value x_j that meets certain criteria, we perform a Bernoulli trial; a failure means the value is kept, while a success means it is sampled as an artificial missing value. The specific success probability p is defined as follows, where μ_j and σ_j are the mean and standard deviation of the j -th column in the dataset, respectively:

- If a value $x_j < \mu_j - 3\sigma_j$, the success probability p of it being kept in the Bernoulli trial is 0.1.
- If $\mu_j - 3\sigma_j \leq x_j < \mu_j - 2\sigma_j$, the success probability p is 0.4.
- If $\mu_j - 2\sigma_j \leq x_j < \mu_j - \sigma_j$, the success probability p is 0.8.

5.2.4 Loss Function

The total loss function of PRISM, L_{total} , is composed of two parts: a reconstruction loss, L_{MSE} and a missingness prediction loss, L_{CE} .

Reconstruction Loss (L_{recon}) We use the Mean Squared Error to measure the difference between the original sample x and the reconstructed sample $\hat{x}$. Importantly, this loss is calculated only at the positions that were artificially

masked (i.e., where $m_j^{(\text{art})} = 1$). Its formula is:

$$L_{\text{recon}} = \frac{1}{\sum_j m_j^{(\text{art})}} \sum_{j=1}^N m_j^{(\text{art})} (x_j - \hat{x}_j)^2$$

Missingness Prediction Loss (L_{mask}) We use the Cross-Entropy Loss comparing original loss and predicted mask logits, $\hat{m}^{(\text{art})}$, with the ground truth artificial mask, $m^{(\text{orig})}$.

$$L_{\text{mask}} = \text{CrossEntropyLoss}(\hat{m}, m^{(\text{orig})})$$

Total Loss (L_{total}) The total loss is the weighted sum of the two losses, balanced by a hyperparameter λ .

$$L_{\text{total}} = L_{\text{recon}} + \lambda \cdot L_{\text{mask}}$$

5.2.5 Imputation Methods Used For Comparison

To comprehensively evaluate PRISM, we compare it against a broad, representative selection of methods from the literature that are commonly used for imputation. These methods can be grouped into the following categories:

i Single-value methods These methods either impute missing abundance values with the same single value or, for each missing value, draw a value from a predetermined simple distribution. We evaluated Minimum value imputation, where missing values in each feature column are replaced with the observed minimum value of that column, as a simple and commonly used baseline.

ii Local similarity methods These methods assume that missing values of a protein can be predicted from the abundance values of other proteins with similar expression profiles. We selected K-Nearest Neighbors (KNN) imputation as it is simple and widely used, for which we utilized the implementation from the scikit-learn library. We also included an approach based on Multivariate Imputation by Chained Equations (MICE), which iteratively fills in missing data through a series of regression models.

iii Global structure methods These methods assume that proteomics datasets contain redundant information and can, thus, be well described by a low-dimensional representation, which is leveraged for inferring missing values. We used a custom implementation of an algorithm based on Non-negative Matrix Factorization (NMF). To include a comparison against current deep learning (DL) methods, we also considered the recently proposed PIMMS framework.

6 Data Availability

All datasets were downloaded from the links provided in their original publications.

7 Code Availability

The code used for the analysis and experiments in this study is available at <https://github.com/guomics-lab/PRISM>.

8 Acknowledgements

This section acknowledges individuals, organizations, or funding sources that contributed to the study but did not meet the criteria for authorship. It may include funding agencies, collaborators, or institutions that provided support.

9 Author Contributions

T.G directed and supported the study. Z.L designed and implemented the framework and experiments. Y.C contributed to the design of the framework, provided methodological guidance. Z.Y designed and implemented the synthetic data generation. The manuscript was written by Z.L, with all authors involved in reviewing, revising, and approving the final version.

10 Competing Interests

T.G. is a shareholder of Westlake Omics Inc. The remaining authors declare no relevant competing interests.

References

- [1] Tiannan Guo, Judith A Steen, and Matthias Mann. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature*, 638(8052):901–911, 2025.
- [2] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6630–6650, 2022.
- [3] Lisa M Bramer, Jan Irvahn, Paul D Piehowski, Karin D Rodland, and Bobbie-Jo M Webb-Robertson. A review of imputation strategies for isobaric labeling-based shotgun proteomics. *Journal of Proteome Research*, 20(1):1–13, 2020.
- [4] Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods*, 13(9):731–740, 2016.
- [5] Alejandro Brenes, Jens Hukelmann, Dalila Bensaddek, and Angus I Lamond. Multibatch tmt reveals false positives, batch effects and missing values. *Molecular & Cellular Proteomics*, 18(10):1967–1980, 2019.
- [6] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- [7] Mengbo Li and Gordon K Smyth. Neither random nor censored: estimating intensity-dependent probabilities for missing values in label-free proteomics. *Bioinformatics*, 39(5):btad200, 2023.
- [8] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research*, 15(4):1116–1125, 2016.
- [9] Yannis Schumann, Antonia Gocke, and Julia E Neumann. Computational methods for data integration and imputation of missing values in omics datasets. *Proteomics*, 25(1-2):e202400100, 2025.
- [10] Tommi Välikangas, Tomi Suomi, and Laura L Elo. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Briefings in bioinformatics*, 19(6):1344–1355, 2018.
- [11] Soroor Hediye-Zadeh, Andrew I Webb, and Melissa J Davis. Msimpute: Estimation of missing peptide intensity data in label-free quantitative mass spectrometry. *Molecular & Cellular Proteomics*, 22(8), 2023.
- [12] Robin Richter, Juliana F Tavares, Anne Miloschewski, Monique MB Breteler, and Sach Mukherjee. Imputebench: Benchmarking single imputation methods. *medRxiv*, pages 2025–02, 2025.
- [13] Bobbie-Jo M Webb-Robertson, Holli K Wiberg, Melissa M Matzke, Joseph N Brown, Jing Wang, Jason E McDermott, Richard D Smith, Karin D Rodland, Thomas O Metz, Joel G Pounds, 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(5):1993–2001, 2015.
- [14] Lincoln Harris, William E Fondrie, Sewoong Oh, and William S Noble. Evaluating proteomics imputation methods with improved criteria. *Journal of proteome research*, 22(11):3427–3438, 2023.

- [15] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
- [16] Alexander Kowarik and Matthias Templ. Imputation with the r package vim. *Journal of statistical software*, 74:1–16, 2016.
- [17] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
- [18] Shisheng Wang, Wenxue Li, Liqiang Hu, Jingqiu Cheng, Hao Yang, and Yansheng Liu. Naguider: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic acids research*, 48(14):e83–e83, 2020.
- [19] Matthias Mann, Chanchal Kumar, Wen-Feng Zeng, and Maximilian T Strauss. Artificial intelligence for proteomics and biomarker discovery. *Cell systems*, 12(8):759–770, 2021.
- [20] Jakob Nybo Nissen, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, Henrik Bjørn Nielsen, Thomas Nordahl Petersen, Ole Winther, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nature biotechnology*, 39(5):555–560, 2021.
- [21] Zhiwei Liu, Pu Liu, Yingying Sun, Zongxiang Nie, Xiaofan Zhang, Yuqi Zhang, Yi Chen, and Tiannan Guo. Dia-bert: pre-trained end-to-end transformer models for enhanced dia proteomics data analysis. *Nature Communications*, 16(1):3530, 2025.
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [23] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, pages 373–382. Springer, 2017.
- [24] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- [25] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.
- [26] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 240–248, 2020.
- [27] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.
- [28] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-miwae: Deep generative modelling with missing not at random data. *arXiv preprint arXiv:2006.12871*, 2020.
- [29] Lincoln Harris and William S Noble. Imputation of cancer proteomics data with a deep model that learns from many datasets. *bioRxiv*, 2024.
- [30] Henry Webel, Lili Niu, Annelaura Bach Nielsen, Marie Locard-Paulet, Matthias Mann, Lars Juhl Jensen, and Simon Rasmussen. Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. *Nature Communications*, 15(1):5405, 2024.
- [31] Tobias Pietz, Sukrit Gupta, Christoph N Schlaffner, Saima Ahmed, Hanno Steen, Bernhard Y Renard, and Katharina Baum. Pepermint: peptide abundance imputation in mass spectrometry-based proteomics using graph neural networks. *Bioinformatics*, 40(Supplement_2):ii70–ii78, 2024.
- [32] Lucas Etourneau, Laura Fancello, Samuel Wiczorek, Nelle Varoquaux, and Thomas Burger. Penalized

444 likelihood optimization for censored missing value imputation in proteomics. *Biostatistics*, 26(1):kxaf006,
445 2025.

446 [33] Katarzyna Woźnica and Przemysław Biecek. Does imputation matter? benchmark for predictive models. *arXiv*
447 *preprint arXiv:2007.02837*, 2020.

448 [34] Henry Webel, Yasset Perez-Riverol, Annelaura Bach Nielsen, and Simon Rasmussen. Mass spectrometry-based
449 proteomics data from thousands of hela control samples. *Scientific Data*, 11(1):112, 2024.

450 [35] Gennifer E Merrihew, Jea Park, Deanna Plubell, Brian C Searle, C Dirk Keene, Eric B Larson, Randall Bateman,
451 Richard J Perrin, Jasmeer P Chhatwal, Martin R Farlow, et al. A peptide-centric quantitative proteomics dataset
452 for the phenotypic assessment of alzheimer’s disease. *Scientific data*, 10(1):206, 2023.

453 [36] Rui Sun, Liujia Qian, Yongge Li, Honghan Cheng, Zhangzhi Xue, Xuedong Zhang, Lingling Tan, Yuecheng
454 Zhan, Wenbin Hu, Qi Xiao, et al. A perturbation proteomics-based foundation model for virtual cell construction.
455 *bioRxiv*, pages 2025–02, 2025.

456 [37] Rui Sun, A Jun, Haolan Yu, Yan Wang, Miaoxia He, Lingling Tan, Honghan Cheng, Jili Zhang, Yingrui Wang,
457 Xiaochen Sun, et al. Proteomic landscape profiling of primary prostate cancer reveals a 16-protein panel for
458 prognosis prediction. *Cell Reports Medicine*, 5(8), 2024.

459 [38] Yaoting Sun, Sathiyamoorthy Selvarajan, Zelin Zang, Wei Liu, Yi Zhu, Hao Zhang, Wanyuan Chen, Hao
460 Chen, Lu Li, Xue Cai, et al. Artificial intelligence defines protein-based classification of thyroid nodules. *Cell*
461 *discovery*, 8(1):85, 2022.

462 [39] Liujia Qian, Jianqing Zhu, Zhangzhi Xue, Yan Zhou, Nan Xiang, Hong Xu, Rui Sun, Wangang Gong, Xue Cai,
463 Lu Sun, et al. Proteomic landscape of epithelial ovarian cancer. *Nature Communications*, 15(1):6462, 2024.

464 [40] Haoran Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using
465 copula functions. In *Advances in database technology: proceedings. International conference on extending*
466 *database technology*, volume 2014, page 475, 2014.

467 [41] Oliver M Crook, Chun-wa Chung, and Charlotte M Deane. Challenges and opportunities for bayesian statistics
468 in proteomics. *Journal of proteome research*, 21(4):849–864, 2022.

469 [42] Peter J Rousseeuw and Geert Molenberghs. Transformation of non positive semidefinite correlation matrices.
470 *Communications in Statistics–Theory and Methods*, 22(4):965–984, 1993.

471 [43] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical*
472 *Association*, 46(253):68–78, 1951.

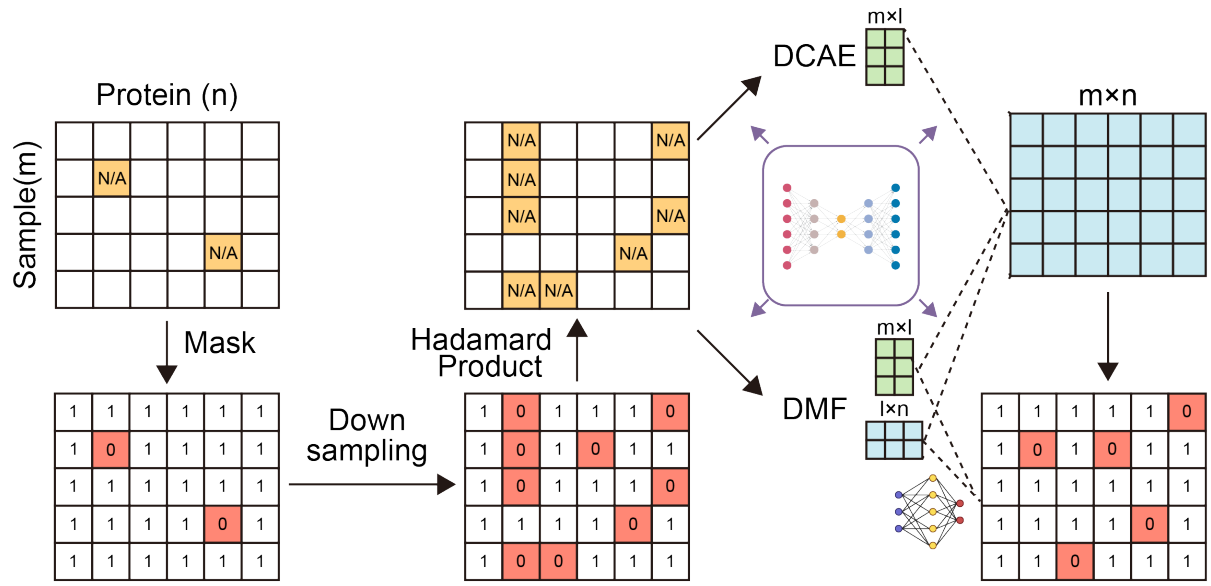


Figure 1: Architecture and workflow of the two core imputation models within the PRISM framework. The framework consists of a model based on Deep Matrix Factorization (DMF, bottom path) and a model based on a Denoising Convolutional Autoencoder (DCAE, top path). In the DMF model, separate embeddings for samples and proteins are combined and fed into a MLP to reconstruct the full matrix. In the DCAE model, an entire sample vector is encoded into a latent space representation via convolutional layers and then decoded back via transposed convolutional layers for reconstruction. Both models employ a dual-task learning strategy, simultaneously reconstructing data and predicting the missingness pattern, and are jointly optimized by a combined loss function.

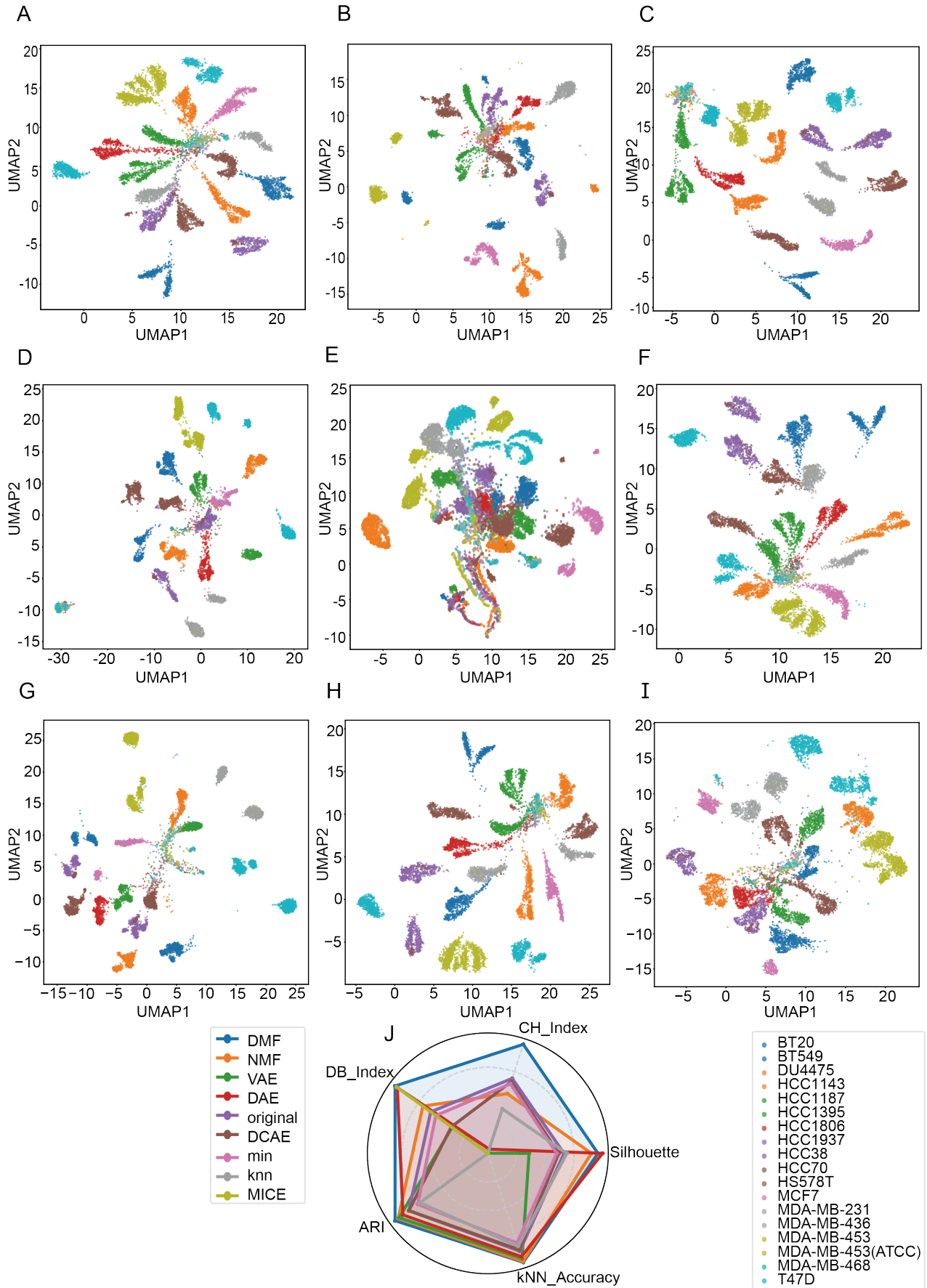


Figure 2: (A-I) UMAP visualization of the imputed data, where each color represents a distinct breast cancer cell line and the separation of the colored clusters indicates the preservation of the underlying biological structure. The panels correspond to data imputed by the following methods: (A) DCAE, (B) DAE, (C) DMF, (D) k-Nearest Neighbors (knn), (E) MICE, (F) Minimum value imputation (min), (G) NMF, (H) original data with zero-filling, and (I) VAE. (J) A radar chart comparing the performance of these methods across five key clustering and manifold learning metrics. The methods in the radar chart legend are sorted in descending order of their Adjusted Rand Index (ARI) score.

Table 1: Statistics of the real-world datasets used in the study.

Dataset	Samples	Proteins	Original Missing Rate (%)
Alzheimer	210	1,541	21.21
HeLa	50	4,535	17.30
PCSHA	918	10,053	44.61
PTV1	16223	5584	65.69
TPD	1,132	6,749	51.00
ZZOV	1,041	10,527	35.78

Table 2: RMSE of DMF model at different mask rates.

Dataset	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Alzheimer	0.0956	0.0953	0.0964	0.0959	0.0974	0.0930	0.0915	0.929	0.0946	0.0966	0.0945
HeLa	0.0880	0.0820	0.0816	0.0820	0.0806	0.0816	0.0821	0.0808	0.0826	0.0833	0.0835
PCSHA	0.0522	0.0525	0.0519	0.0516	0.0523	0.0518	0.0515	0.0515	0.0529	0.0520	0.0532
PTV1	0.1042	0.1016	0.1019	0.1018	0.1006	0.1018	0.1016	0.1031	0.1018	0.1019	0.1015
Synthetic Data	0.1412	0.1235	0.1361	0.1342	0.1330	0.1268	0.1335	0.1335	0.1353	0.1339	0.1378
TPD	0.0285	0.0285	0.0273	0.0268	0.0261	0.0271	0.0270	0.0270	0.0273	0.0272	0.0269
ZZOV	0.0504	0.0503	0.0507	0.0512	0.0500	0.0501	0.0507	0.0491	0.0493	0.0508	0.0597

Table 3: RMSE of DCAE model at different mask rates.

Dataset	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Alzheimer	0.1307	0.1096	0.1035	0.1102	0.1108	0.1098	0.1079	0.1070	0.1088	0.1109	0.1088
HeLa	0.1098	0.0850	0.0833	0.0837	0.0838	0.0833	0.0823	0.0825	0.0843	0.0837	0.0852
PCSHA	0.0815	0.0622	0.0612	0.0609	0.0616	0.0617	0.0608	0.0608	0.0618	0.0614	0.0626
PTV1	0.1038	0.0942	0.0899	0.0892	0.0861	0.0898	0.0902	0.0901	0.0897	0.0941	0.0888
Synthetic Data	0.1632	0.1605	0.1589	0.1655	0.1649	0.1606	0.1681	0.1662	0.1673	0.1660	0.1698
TPD	0.0184	0.0165	0.0157	0.0157	0.0157	0.0157	0.0150	0.0157	0.0157	0.0158	0.0155
ZZOV	0.0599	0.0498	0.0511	0.0506	0.0496	0.0497	0.0501	0.0478	0.0503	0.0489	0.0492

Table 4: RMSE comparison of different methods across datasets.

Dataset	DMF	DCAE	NMF	KNN	MICE	Min	DAE	VAE
Alzheimer	0.0930	0.1098	0.1093	0.1272	0.1183	0.1551	0.1219	0.1121
HeLa	0.0816	0.0833	0.0872	0.0892	0.0933	0.1267	0.1219	0.0992
PCSHA	0.0519	0.0618	0.0514	0.0613	0.0605	0.0845	0.0714	0.0559
PTV1	0.0501	0.1331	0.1266	0.1987	0.1331	0.2173	0.1682	0.1659
Synthetic Data	0.1289	0.1606	0.1359	0.2008	0.1427	0.2144	0.1554	0.1966
TPD	0.0272	0.0157	0.0318	0.0332	0.0458	0.0327	0.0583	0.0324
ZZOV	0.0501	0.0496	0.0596	0.0659	0.0664	0.0738	0.0663	0.0573

Table 5: Classification accuracy comparison between different methods imputed datasets.

Dataset	DMF	DCAE	NMF	KNN	Original	MICE	Min	DAE	VAE
PCSHA	0.6378	0.6349	0.5949	0.5568	0.6080	0.6372	0.5982	0.5987	0.5835
TPD	0.8033	0.8072	0.7898	0.7741	0.8096	0.7807	0.7733	0.7725	0.7865
ZZOV	0.7348	0.7271	0.7064	0.7267	0.7262	0.7228	0.7247	0.7041	0.7258

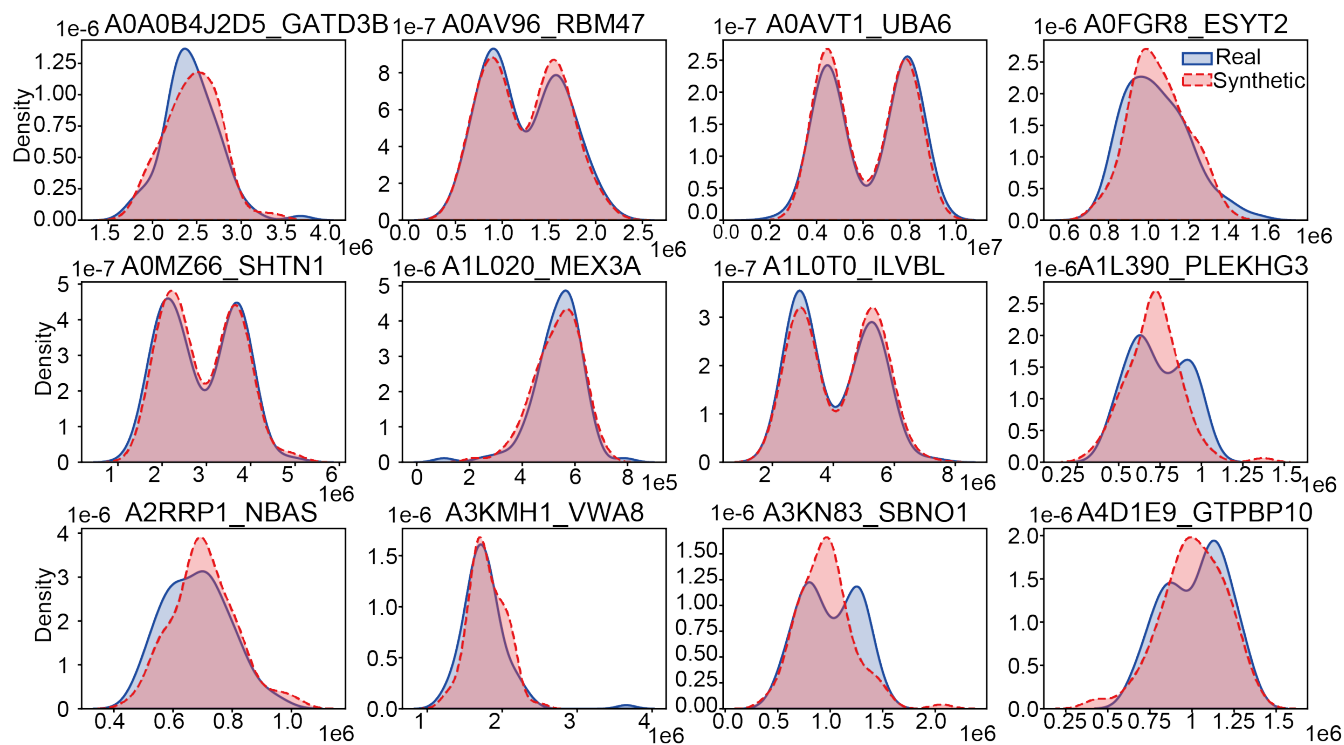


Figure S1: Comparison of marginal distributions for selected proteins in the original (blue) versus synthetic (orange) data.