Penalized likelihood optimization for censored missing value imputation in proteomics

Lucas Etourneau[1,2], Laura Fancello[1], Samuel Wieczorek[1], Nelle Varoquaux[2], and Thomas Burger[1,*]

[1]Univ. Grenoble Alpes, CNRS, CEA, Inserm, BGE UA13, ProFI FR2048, 38000, Grenoble, France
[2]TIMC, Univ. Grenoble Alpes, CNRS, Grenoble INP, 38000, Grenoble, France
[*]thomas.burger@cea.fr

March 18, 2025

Abstract

Label-free bottom-up proteomics using mass spectrometry and liquid chromatography has long been established as one of the most popular high-throughput analysis workflows for proteome characterization. However, it produces data hindered by complex and heterogeneous missing values, which imputation has long remained problematic. To cope with this, we introduce Pirat, an algorithm that harnesses this challenge using an original likelihood maximization strategy. Notably, it models the instrument limit by learning a global censoring mechanism from the data available. Moreover, it estimates the covariance matrix between enzymatic cleavage products (*i.e.*, peptides or precursor ions), while offering a natural way to integrate complementary transcriptomic information when multi-omic assays are available. Our benchmarking on several datasets covering a variety of experimental designs (number of samples, acquisition mode, missingness patterns, etc.) and using a variety of metrics (differential analysis ground truth or imputation errors) shows that Pirat outperforms all pre-existing imputation methods. Beyond the interest of Pirat as an imputation tool, these results pinpoint the need for a paradigm change in proteomics imputation, as most pre-existing strategies could be boosted by incorporating similar models to account for the instrument censorship or for the correlation structures, either grounded to the analytical pipeline or arising from a multi-omic approach.

Keywords: (1) Mass Spectrometry Based Proteomics; (2) Imputation of Missing Not At Random Values; (3) Multi-Omic Imputation; (4) Covariance Matrix Estimation; (5) Penalized Likelihood Maximization

1 Introduction

Bottom-up label-free LC-MS/MS (tandem mass spectrometry coupled with liquid chromatography) stands out as the most widely used method to characterize the proteome of biological samples. These assays typically yield an abundance matrix, where each row corresponds to a sample, each column a protein. Yet, moving from the massive amount of MS signals generated to such abundance matrix requires robust, mathematically well-grounded, and fast evolving software, posing computational and theoretical challenges. Notably two difficulties specific to bottom-up workflows arise. First, protein abundances are not directly measured: Sample preparation involves proteolysis (*i.e.*, the cleavage of each protein into several *peptides* using an enzyme like trypsin) and MS measurements require the peptides to be ionized beforehand (to which we refer to as *precursor ions*, or simply *precursors*). Thus, inferring the identity and quantity of proteins requires two aggregation steps: from precursors to peptides, and from peptides to proteins. Second, raw LC-MS/MS data are hampered by a large number of missing values (MVs) in the abundance matrices: the overall MVs rate can reach 50% [Lazar et al., 2016]—whether that is at the level of precursor, peptide, or protein abundances; and at least 50% of peptides usually have at least one MV [Liu and Dongre, 2021].

The origin of MVs is complex and multi-faceted. Since the work of Karpievitch et al. [2012], it is customary to separate MVs into two types. The first one gathers all the MVs resulting from the various workflow imperfections (such as peptide ionization issues, enzymatic miscleavages, too complex samples, false peptide identifications, *etc.*) and which may affect proteins broadly randomly, regardless of their abundance. The second category corresponds to the censorship mechanism resulting from the instrument limit: precursors with an abundance below this limit yield MVs. A major issue with this censorship relates to its stochastic and dynamic nature, as the MS range changes along the acquisition process [Vidova and Spacil, 2017]. As a downside of the dramatic proportion and complex nature of MVs, analysing only fully observed biomolecules (be they precursors, peptides or proteins) is usually not considered, as it would result in discarding too much biologically relevant information.

While it is theoretically safer to conduct the analysis without inferring MVs [Little and Rubin, 2019], *e.g.*, identifying differentially expressed proteins using models that cope with MVs [Chen et al., 2014, Ryu et al.,

2014, O'brien et al., 2018, Goeminne et al., 2020, Chion and Leroy, 2023], many downstream investigation techniques require in practice to impute first MVs (*i.e.*, to estimate the missing abundances). Yet, despite extensive research, this remains a fundamental problem of computational proteomic research.

Drawing from a vast body of literature that spans a decade, some consensus on the subject has emerged. First, imputing missing protein abundances is suboptimal [Lazar et al., 2016] as a result of the effect of the peptide-to-protein aggregation on missing values. However, whether imputation should be performed at precursor- or peptide-level has not been investigated yet. Second, it is insightful to distinguish MVs according to the following well-acknowledged statistical categories [Little and Rubin, 2019]: Missing Completely At Random (MCAR), where the probability that a value is missing does not depend on any observed or missing value in the data; Missing At Random (MAR), where the same probability may only depend on observed values; and Missing Not At Random (MNAR) in any other case. Accordingly, in bottom-up label free proteomics, MVs resulting from the lower instrumental limit are classically assumed to follow an MNAR mechanism [Karpievitch et al., 2012, Webb-Robertson et al., 2015, Lazar et al., 2016], while other MVs, in absence of sufficiently refined MAR-based description, are modelled as MCARs. Third, the algorithms providing the best imputations are not the same depending on whether MVs are classified as MCARs or MNARs [Lazar et al., 2016]. This observation has motivated the development of meta-imputation tools, *i.e.*, algorithms taking as input one or several imputation methods(s), and delivering as output a more refined imputation result *e.g.*, [Wei et al., 2018, Ma et al., 2020, Giai Gianetto et al., 2020, Gardner and Freitas, 2021, Wang et al., 2022, Chion et al., 2022], as well as diagnosis tools capable of proposing an imputation strategy tailored to the data specificities [Wang et al., 2020, Kong et al., 2022]. Regardless of their practical interest, leveraging those approaches is only possible if classical imputation methods are available in the first place.

Numerous exhaustive reviews propose comparisons between a wide range of imputation algorithms, see for example [Lazar et al., 2016, Karpievitch et al., 2012, Webb-Robertson et al., 2015, Jin et al., 2021, Liu and Dongre, 2021, Shen et al., 2022], as well as Section 3.2. Although those works generally concur on the least accurate approaches, they do not agree on the most accurate ones, which depend on the dataset or on the experimental design. Moreover, while MCAR/MAR-devoted methods have established their robustness in many scientific domains where MVs have hardly specific behaviors, the MNAR-devoted methods used in proteomics are often simple and univariate, thus poorly informative [Chion and Leroy, 2023] and leading to larger imputation errors. A notable exception is msImpute [Hediyeh-zadeh et al., 2023], whose very recent publication has unveiled promising results. Contrarily to most anterior methods, msImpute simultaneously tackle MCARs and MNARs, by estimating the type of each MV and interpolating MAR and MNAR imputation distributions accordingly. Yet, some parameters (as the barycenter weights or the MNAR assumption) are fixed in a predetermined manner, which may hinder the generalization capabilities of the approach. Pushing the logic a step further, few works [Chen et al., 2014, Li and Smyth, 2023] propose to bypass any MCAR/MNAR distinction, assuming the relevant model should directly estimate the probability of missing depending on the abundance. Among them, PEMM [Chen et al., 2014] has been an important source of inspiration for this work. Considering an analytical approximation of a left-censoring mechanism, it uses the Expectation-Maximization algorithm to maximize a penalized likelihood model of the mean and covariance matrix of the protein-level data, yielding natural parametric imputations. Unfortunately, this essentially theoretical proposal has shown many practical limitations that we detail in Sup. Mat. A.1, notably important convergence issues, which explains its scarce use on real proteomics data.

More broadly, proteomic MV imputation aims at filling gaps in otherwise difficult to interpret biological data: So does multi-omic data integration. This field has elaborated on the natural assumption that a single omic modality contains only partial information which can be compensated for by applying multi-omics technologies to the same or related samples. To do so, one often estimates a common latent distribution modelling the underlying biological phenomena—classically using either matrix factorization [Argelaguet et al., 2018, Leppäaho et al., 2017, Rohart et al., 2017, Meng et al., 2019] or deep learning [Zhou et al., 2020, Barzine et al., 2020]). Then, the latent model is leveraged in a variety of tasks, such as pathway identification [Rohart et al., 2017, Argelaguet et al., 2018, Meng et al., 2019], sample extrapolation [Rohart et al., 2017, Barzine et al., 2020, Zhou et al., 2020], gene-set analysis [Meng et al., 2019], and possibly imputation [Flores et al., 2023]. It is worth noting that none of the available methods was specifically developed (and tested) to fulfil this task in the challenging context of bottom-up label-free LC-MS/MS data. We have however noticed a recent (still unpublished) attempt by Gupta et al. [2023] to train graph neural network in a gene-specific manner to impute protein-level MVs from such data, following a logic broadly akin to the one presented here.

In this work, we propose a concrete route to improve the imputation of LC-MS/MS-based label-free bottom-up proteomic data by leveraging a well-established analytical and biochemical truism: precursors or peptides originating from the same protein should exhibit correlations that are insightful for proteomic MV imputation. Using this strategy, it becomes possible to accurately impute even with only few observed values per peptide, making our algorithm dramatically more performing than state of the art methods in the classical proteomic context of scarce samples. In addition, as insightful correlations can also be exploited between different omics modalities, we have generalized the approach to incorporate quantitative transcriptomic data, hereby

opening the path to multi-omic based imputation of proteomic data. Doing so is however not sufficient to fully tackle the MNAR issue. Therefore, we have implemented these concepts using an imputation algorithm which estimates a single model for both censored and random MVs. Although our censorship model roots on that of PEMM [Chen et al., 2014], we propose a novel estimation strategy, more stable, scalable, and, unlike PEMM, without parameter-tuning or functional approximations, as to fit the practical constraints of proteomic data analysts. The resulting software, referred to as Pirat and released on the Bioconductor (`https://www.bioconductor.org/packagesT7SS-ESX1-EccB-PF05108/release/bioc/html/Pirat.html`), significantly outperforms all other imputation methods on a variety of tasks, datasets, and situations.

The article is organized as follows: The proposed method is described in Section 2. Then, sections 3 and 4 respectively present the validation protocol and results. Finally the paper ends with a discussion (Section 5) and conclusions (Section 6). All our results are reproducible with the code available on GitHub (`https://github.com/TrEE-TIMC/Pirat_experiments`).

2 Method

2.1 Notations

We hereafter use the following notations:

- X: a complete matrix (of size $n \times p$) of peptide $\log_2$ abundances, with rows referring to samples and columns to peptides. Note that depending on the context, X may refer to the entire dataset, or to a subset of it, namely a set of peptides or of precursor ions having the same mother protein(s).
- $X_{i,j}$: the abundance value in X of the i-th row (sample i) and j-th column (peptide j); when i (respectively, j) is replaced by "." one refers to the to j-th column of X (respectively, the i-th column of X): $X_{.,j}$ thus depicts the peptide vector (respectively, $X_{i,.}$ depicts the sample vector).
- M: an indicator matrix of the same size as X, reflecting whether an abundance value is missing or not. The same indexing notations as of X applies to M ($M_{i,j}$, $M_{.,j}$, $M_{i,.}$).
- $M_{i,j}$: missingness indicator of abundance value of j-th peptide in i-th sample (1 if missing, 0 otherwise).
- $X_{\text{obs}}, M_{\text{obs}}$ (resp. $X_{\text{mis}}, M_{\text{mis}}$): all values of X and M corresponding to observed (resp. missing) abundance values.
- x, m, *etc.*: We use uppercase letters (*e.g.*, $X_{i,j}$ and $M_{i,j}$) when considering random variables and lowercase ones (*e.g.*, $x_{i,j}$ and $m_{i,j}$) when considering their realisations.
- $\{i, \text{obs}\}$ (resp. $\{i, \text{mis}\}$): for a given sample i, set of (i, j) pairs such that $x_{i,j}$ is observed (resp. missing). For clarity, brackets are removed when using this notation as subscript.
- μ: the vector of the theoretical mean abundance of the p peptides.
- Σ: the covariance matrix of the peptide's abundances, of size $p \times p$. The same indexing notation as for matrices X and M ($\Sigma_{j,j'}$ indicates the covariance between peptides j and j').
- Γ: the set of parameters defining the missingness mechanism (see definition in Section 2.3).
- I denotes the identity matrix.

2.2 A bird's eye view on Pirat methodology

Our method Pirat (standing for *Precursor or Peptide Imputation under Random Truncation*) works as follows: Firstly, depending on the input data, it creates either Peptide Groups or Precursor Groups. Both are based on protein belonging and are abbreviated as PGs, whereas peptides or precursors in a same PG are qualified as *siblings*. For sake of readability, we hereafter refer to peptides only, precursors being mentioned only if they require specific processing. It is usually assumed that sibling peptides will quantitatively behave similarly across treatment groups/biological conditions, even though the presence of isoforms, post-translational modifications, or miscleavage can impact these correlations [Dermit and Meyer, 2021]. Peptides from nested proteins form a unique PG and we duplicate peptides that are shared between PGs. The resulting biochemically informed dependence graph can optionally be enriched with transcriptomic data: for any PG, one then appends the sample-wise abundance vector(s) of related transcript(s).

Secondly, Pirat estimates over all the peptides a global missingness mechanism:

$$P(M_{i,j} = 1 | X_{i,j}, \Gamma) = \min\left(1, \exp(-\gamma_0 - \gamma_1 X_{i,j})\right) \ , \tag{1}$$

The missingness parameter $\Gamma = (\gamma_0, \gamma_1)$ is estimated using a linera regression on a sliding window as to fit the missingness pattern of each dataset.

Thirdly, Pirat estimates the same model as PEMM [Chen et al., 2014], which is derived from the model of Little and Rubin [2019], yet using an original approach. Briefly, it aims at maximizing the penalized log-likelihood of observed data and missingness response, which, considering Γ known, amounts to maximizing the following quantity:

$$\log P(X_{\text{obs}}|\mu, \Sigma) + \log \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\mu,\Gamma} \left[P(M_{\text{mis}}|X_{\text{mis}}, \Gamma)\right] + Q_{K,\lambda}(\Sigma) \ ,$$

where Q is a penalty on Σ. In Pirat, we fit this model independently on each PG, with respect to μ and Σ. Likewise, penalty hyperparameters K and λ are automatically set for each PG in the empirical Bayes framework. To avoid the parametric approximations of the missingness mechanism that hamper PEMM's results, we propose a tractable and differentiable lower bound of the above quantity. Therefore, the lower bound can be maximized with respect to μ and Σ using L-BFGS, a quasi-Newton optimization procedure proposed by Liu and Nocedal [1989]. To decrease computation costs and memory usage, as well as to enforce Σ positive definiteness, Pirat leverages an alternative parameterization of Σ based on its log-Cholesky factorization.

Fourthly, once the model is fitted, we propose to impute the MVs by their conditional mean with respect to observed values and their missingness response. The conditional mean being devoid of closed form, it is computed using standard Monte-Carlo integration [Robert and Casella, 1999]. The multiple imputed values obtained for shared peptides are averaged.

2.3 Model's assumptions

Our model relies on the following four main assumptions regarding the whole peptide dataset (here, X and M refer to all the peptides of all the PGs of the dataset):

1. The $X_{i,.}$ are i.i.d. and normally distributed with parameters μ and Σ.
2. The missingness mechanism is self-masked for all peptides [Sportisse et al., 2020], *i.e.*, the probability of an abundance value being missing, knowing the entire sample's abundances, only depends on the abundance itself:
$$\forall i, j \quad P(M_{i,j} = 1|X_{i,.}) = P(M_{i,j} = 1|X_{i,j}); \tag{2}$$
Or equivalently,
$$\forall i, j \quad M_{i,j} \perp\!\!\!\perp X_{i,-j}|X_{i,j} \text{ where } -j \text{ refers to the set } [\![0, p]\!] \setminus \{j\}. \tag{3}$$
3. The missingness responses for each peptide of a sample are independent conditionally to their abundance [Sportisse et al., 2020]:
$$\forall i, j, k \text{ with } k \neq j, \quad M_{i,j} \perp\!\!\!\perp M_{i,k}|X_{i,.} \tag{4}$$
4. The missingness mechanism can be written as in [Chen et al., 2014] :
$$P(M_{i,j} = 1|x_{i,j}, \Gamma) = \min(1, \exp(-\gamma_0 - \gamma_1 x_{i,j})), \text{ where } \Gamma = \{\gamma_0, \gamma_1\} \text{ with } \gamma_1 \geq 0; \tag{5}$$
We assume here that there is a minimum detection threshold $-\gamma_0/\gamma_1$ below which no peptide can be quantified, and that the probability for a peptide to be missing exponentially decays with its log-abundance.

These assumptions ensure identifiability of μ, Σ, and Γ parameters in univariate and multivariate cases, as demonstrated by Sportisse et al. [2020] and by Miao et al. [2016].

2.4 Estimation of the missingness parameters

The set of parameters Γ over the whole peptide dataset is estimated once and for all as follows:

1. Sort the peptides by their observable mean, and denote by α_i the mean of the i-th peptide i in the ordered vector α;
2. For $i = \lceil \frac{k+1}{2} \rceil, \ldots, p + 1 - \lfloor \frac{k+1}{2} \rfloor$ (where $\lceil . \rceil$ and $\lfloor . \rfloor$ denotes the upper and lower rounding), compute the rolling average of the missingness percentage y_i over the ordered peptides, with a window of size k; which leads to $p - k + 1$ points (α_i, y_i);
3. Fit a linear model on the points (α_i, y_i) by ordinary least squares, and set the parameters of the missingness mechanism (see Equation 5) as the coefficients obtained.

Using $k = 10$, preliminary experiments have shown a clear and smooth decreasing trend of $\log(y)$ with respect to α and a satisfying goodness of fit, as illustrated in Sup. Mat. Figure 11. Therefore, this value is used in all further experiments.

2.5 Penalized likelihood model

Considering the missingness parameters Γ known, we now propose to follow the selection model of Little and Rubin [2019] and to maximize the joint log-likelihood $\mathcal{L}$ of the observed values and of the missingness response with respect to the Gaussian parameters μ and Σ, iteratively for each PG:

$$(\hat{\mu}, \hat{\Sigma}) = \underset{\mu,\Sigma}{\operatorname{argmax}} \quad \mathcal{L}(\mu, \Sigma | X_{\text{obs}}, M, \Gamma) + Q_{K,\lambda}(\Sigma)\,, \tag{6}$$

where X and M hereafter refer to the abundance matrix and missingness response of a single PG, and where $Q_{K,\lambda}(\Sigma)$ is a penalty term on Σ with hyperparameters K and λ (see Section 2.6), as proposed by Chen et al. [2014]. Hence, maximizing this penalized log-likelihood amounts to maximizing the posterior distribution of parameters μ, Σ, and Γ. Considering previous assumptions, the log-likelihood of the selection model decomposes as:

$$\begin{aligned}
\mathcal{L}(\mu, \Sigma | X_{\text{obs}}, M, \Gamma) &= \log P(X_{\text{obs}} | \mu, \Sigma) + \log P(M | X_{\text{obs}}, \mu, \Sigma, \Gamma) \text{ (Bayes formula)} && (7)\\
&= \log P(X_{\text{obs}} | \mu, \Sigma) + \log \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\mu,\Sigma} \left[P(M | X_{\text{obs}}, X_{\text{mis}}, \Gamma) \right] \\
&= \log P(X_{\text{obs}} | \mu, \Sigma) + \log \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\mu,\Sigma} \left[P(M_{\text{obs}} | X_{\text{obs}}, \Gamma) P(M_{\text{mis}} | X_{\text{mis}}, \Gamma) \right] \text{ (As. 2 \& 3)} \\
&= \log P(X_{\text{obs}} | \mu, \Sigma) + \log P(M_{\text{obs}} | X_{\text{obs}}, \Gamma) + \log \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\mu,\Sigma} \left[P(M_{\text{mis}} | X_{\text{mis}}, \Gamma) \right] \\
\mathcal{L}(\mu, \Sigma | X_{\text{obs}}, M, \Gamma) &= \log P(X_{\text{obs}} | \mu, \Sigma) + \log P(M_{\text{obs}} | X_{\text{obs}}, \Gamma) && (8)\\
&\quad + \sum_{i}^{n} \log \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma} \left[\prod_{j \in \{i,\text{mis}\}} P(M_{i,j} = 1 | X_{i,j}, \Gamma) \right] .
\end{aligned}$$

Note that the term $\log P(M_{\text{obs}} | X_{\text{obs}}, \Gamma)$ does not depend on μ and Σ. Thus, if we consider Γ known, we can remove this term from the log-likelihood to optimize the parameters.

2.6 Penalty over Σ

We estimate Σ and μ for each PG independently by optimizing Equation 6. Therefore, the penalty hyperparameters λ and K must be tuned automatically (as many times as PGs). To do so, we rely on a Bayesian interpretation of the penalty term $Q_{K,\lambda}(\Sigma)$: it can be viewed as an inverse Wishart prior of the covariance matrix Σ. This penalty can be rewritten as:

$$\begin{aligned}
Q_{K,\lambda}(\Sigma) &= \frac{1}{2} 2K \log(|\Sigma|) + \frac{1}{2} tr(2\lambda I \Sigma^{-1}) && (9)\\
&= \log[p_{\mathcal{W}^{-1}(2K-p-1, 2\lambda I)}(\Sigma)] + C\,,
\end{aligned}$$

where $p_{\mathcal{W}^{-1}(2K-p-1,2\lambda I)}$ denotes the density of an inverse Wishart distribution with parameters $(2K-p-1, 2\lambda I)$, and where C is constant with respect to Σ. Hence, K is related to the number of degrees of freedom in Σ estimate, and λ is related to the scaling matrix $2\lambda I$.

This leads us to empirically estimate the hyperparameters of each PG from the set of all fully observed peptides by relying on the empirical Bayes framework of Efron and Morris [1972] where we leverage that, in the univariate case, the inverse Wishart distribution boils down to the inverse-gamma one. Specifically, we compute their empirical variance and estimate the parameters of an inverse-gamma distribution (denoted α and β) through an MLE. Finally, by setting $K = \alpha + p$ and $\lambda = \beta$, the marginal prior distribution of each diagonal element of Σ (the variance of each peptide) is an inverse-gamma with parameters (α, β). Hence, the peptide's variances have the same prior distribution among all PGs (this property is instrumental as before looking at any data, one would not expect the variance of a peptide to depend on the PG it belongs to). Without relaxing this property, we increase α by a factor 2 to better constrain the estimation of the covariance matrix (following Chen et al. [2014], which showed in various scenarios that increasing K improves the estimation of Σ).

2.7 Deriving a lower-bound of the penalized log-likelihood to estimate μ and Σ

A main issue regarding the estimation of μ and Σ by maximizing Equation 7, is that the expectation inside the log-likelihood is not analytically tractable for any missingness mechanism. The authors of PEMM proposed an approximation of the missingness mechanism that profoundly hinders the convergence of their model, as we show in Sup. Mat. A.1 with experiments on both real and synthetic data.

Another approach consists in using Jensen's inequality, and the original missingness mechanism from Equation 5. We can then derive a tractable and differentiable lower bound of the likelihood that can be optimized. Concretely, to estimate μ and Σ for a given PG, we propose to maximize the following lower bound,

$\forall i \in \{1, ..., n\}$:

$$\log \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)] \geq \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}\Big[\sum_{j\in\{i,\text{mis}\}} \log P(m_{i,j}=1|x_{i,j},\Gamma)\Big]$$
$$\geq \sum_{j\in\{i,\text{mis}\}} \mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big]\,. \tag{10}$$

Let $W_i \sim X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma \sim \mathcal{N}(\tilde{\mu}_i,\tilde{\Sigma}_i)$ and w_i the realisation of W_i. Using the parametric model from Equation 5, $\forall j \in \{i, \text{mis}\}$:

$$\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big] = \int_{-\frac{\gamma_0}{\gamma_1}}^{+\infty} (-\gamma_0-\gamma_1 w_{i,j})\phi(w_{i,j}|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j})\,\mathrm{d}w_{i,j} \tag{11}$$
$$= -\gamma_0\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right) -$$
$$\gamma_1\int_{-\frac{\gamma_0}{\gamma_1}}^{+\infty} w_{i,j}\phi(w_{i,j}|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j})\,\mathrm{d}w_{i,j}\,,$$

where $\phi(\cdot|\mu,\sigma)$ (resp. $\Phi(\cdot|\mu,\sigma)$) denotes the (resp. cumulative) distribution function of a normal distribution with parameters μ,σ. Then, using properties of truncated normal distribution, we have:

$$\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big] = -\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right)\left(\gamma_0+\gamma_1\mathbb{E}\left[W_{i,j}\Big|W_{i,j}\geq-\frac{\gamma_0}{\gamma_1}\right]\right) \tag{12}$$
$$= -\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right)(\gamma_0+\gamma_1\tilde{\mu}_{i,j}) -$$
$$\gamma_1\tilde{\sigma}_{i,j}^2\phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right).$$

The expression of Equation 12 is differentiable with respect to parameters $\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}$. These two are the parameters of the conditional distribution of $X_{i,j}$ with respect to $X_{i,\text{obs}}$. They are obtained by the linear combination of μ and the Schur complement of the covariance matrix of observed values in i-th sample. Schur complement is differentiable with respect to Σ so that $\tilde{\mu}_{i,j}$ and $\tilde{\sigma}_{i,j}$ are differentiable with respect to Σ and μ. Hence, we can use any automatic differentiation tool, combined with optimization algorithm, to maximize the following lower bound of Equation 6:

$$\tilde{\mathcal{L}}(\mu,\Sigma,|X_{\text{obs}},M,\Gamma) = \log P(X_{\text{obs}}|\mu,\Sigma) + \log P(M_{\text{obs}}|X_{\text{obs}},\Gamma) + \tag{13}$$
$$\sum_i^n \sum_{j\in\{i,\text{mis}\}} \mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|X_{i,j},\Gamma)\big] + Q_{K,\lambda}(\Sigma)\,.$$

Specifically, we use L-BFGS (a quasi-newton method) combined with Armijo backtracking line-search, implemented with Pytorch [Shi and Mudigere, 2020].

To ensure Σ positive-definiteness during the optimization process, we re-parametrize Σ by its log-Cholesky factorization [Pinheiro and Bates, 1996]. Finally, to avoid ill-conditioned matrix that could cause numerical instabilities, we apply Tikhonov regularization method and subsequently add a fixed value $\epsilon_\Sigma = 10^{-4}$ to the diagonal of Σ. Doing so does not impact on the optimum solution, as the minimum variance observed in all the datasets processed so far has been $\geq 10^{-3}$.

2.8 Imputation

Once all parameters μ, Σ, and Γ are estimated, we impute missing values by their conditional mean. To do so, we use the following result of Chen et al. [2014]:

$$\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},M_{i,.},\mu,\Sigma,\Gamma}[X_{i,\text{mis}}] = \frac{\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)X_{i,\text{mis}}]}{\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)]} \tag{14}$$

The distribution on which the left expectation is computed is not accessible, however, the distributions of the right expectations are Gaussian. Hence, we use Monte-Carlo integration to estimate them and compute the conditional mean of $X_{i,\text{mis}}$. Note that, using the same approach, the order 2 moment of $X_{i,\text{mis}}$ can be computed by simply squaring it in the above equation, which gives access the variance of $X_{i,\text{mis}}$.

2.9 Imputation in singleton PGs

Singleton PGs (that correspond to proteins with a single and exclusive peptide or precursor ion being observed through MS analysis) regularly occur in proteomics datasets. Those PGs may be difficult to handle with Pirat, as no correlation can be leveraged. To cope with this we propose 3 altenatives: Pirat-2 (where singleton PGs are simply discarded), Pirat-S (standing for "Sample-correlations") and Pirat-T (standing for "Transcriptome-informed"). Pirat-S and Pirat-T work as follows:

2.9.1 Pirat-S

Pirat is based on the assumption the samples are independently distributed, see Section 2.3. Yet, other algorithms like ImpSec [Verboven et al., 2007] assume differently and leverage dependencies between samples. Following the same idea, we propose in Pirat-S to leverage sample-wise correlations to impute peptides of singleton PGs. To do so, we first estimate missingness parameters Γ and hyperparameters K and λ on the whole dataset and impute only PGs of size > 1 with the classical version of Pirat, described above. Second, we compute the empirical sample mean and sample-wise covariance matrix of the transposed imputed part of the dataset. Finally, we impute the remaining MVs by their conditional mean with respect to observed values, only using the observed values and the empirical sample mean and covariance matrix.

2.9.2 Pirat-T

Pirat-T extends Pirat by enabling the integration of transcriptomic quantitative information, when available, as to guide peptide imputation. To do so, it requires a dataset of $\log_2$ mRNA expressions of samples from the same phenotype(s) as the proteomic ones. However, increments in imputation performances may require paired transcriptomic and proteomic samples. It also requires a correspondence table between transcripts and PGs, for instance, based on their original gene(s). In this work, transcriptomics is essentially used to improve singleton PG imputation (see Section 4.3). However, Pirat's methodology is more versatile and the packaged code makes it possible to tune which PGs are imputed using complementary transcriptomic information, depending on their size.

Concretely, Pirat-T works as follows: for each PG, it integrates all the transcripts (i) for which non-zero values are measured in at least two conditions, and (ii) which are associated to at least one of the proteins of the PG. In the case proteomic and transcriptomic samples are paired, the log-count vectors of the transcripts are simply appended to the PGs. Otherwise, a condition-wise mean log-count vector is used instead. *In fine*, each mRNA log-count vector is processed like an additional (fully observed) peptide, and thus contributes to imputation depending on its correlation with the sibling peptide(s).

3 Validation protocol

3.1 Datasets

To conduct our experiments, we rely on the several publicly available datasets: Cox2014 [Cox et al., 2014], Bouysssie2020 [Bouyssié et al., 2020], Huang2020 [Huang et al., 2020], Capizzi2022 [Capizzi et al., 2022], Vilalllongue2022 [Vilallongue et al., 2022], Habowski2020 [Habowski et al., 2020], and Ropers2021 [Ropers et al., 2021]. A summarized description of how these datasets have been constituted is available in Supp. Mat. A.2. On all datasets, we remove peptides or precursors having strictly less than 2 observations. Finally, we apply $\log_2$ transformation both on peptide or precursor intensities and transcript normalized counts (when available).

3.2 List of competitive methods

To benchmark Pirat, we use 15 state-of-the-art imputation methods, many of which have already been pinpointed by NAguideR evaluation software [Wang et al., 2020]: Bayesian Principal Component Analysis (BPCA) [Oba et al., 2003] and Singular Value Decomposition (SVD) [Troyanskaya et al., 2001] parametrized with a number of components equal to the number of compared conditions in each dataset; GMS [Li et al., 2020] with 3 fold cross-validation for the tuning of its inner parameter (termed λ); K-nearest neighbors (KNN) [Troyanskaya et al., 2001] SeqKNN [Kim et al., 2004], trKNN [Shah et al., 2017] with the number of neighbors k set to 10; ImpSeq [Verboven et al., 2007] and its outlier oriented version ImpSeqRob [Branden and Verboven, 2009], Local Least Square (LLS) [Kim et al., 2005], MinProb [Lazar et al., 2016], MissForest [Stekhoven and Bühlmann, 2012], MLE [Love et al., 2014], msImpute in MAR and MNAR versions [Hediyeh-zadeh et al., 2023], and QRILC [Lazar et al., 2016] with default parameters. MissForest and LLS were tested with an input format where peptides are in columns and samples in rows, contrarily to all other methods. Amongst those methods, the following ones are specifically devoted to low abundance censored values: MinProb, msImpute_mnar, QRILC, TrKNN.

While most methods require only 2 observed values (sometimes less) per peptide, GMS and trKNN require at least 3 of them, and both msImpute versions, 4 of them. This is an issue for two reasons: First, because, on controlled datasets, the biological variability is so small that not imputing those peptides before testing for differential abundance does not hamper the results, hereby artificially boosting the feature selection performances with respect to real-life cases. Second, because observing only two values is common for peptides nearby the detection limit, so that a concrete assessment on the MNAR imputation error is not possible with these algorithms. This is why, regardless of their performances on the biomarker selection task (see Section 4.1), GMS, trKNN, msImpute_MNAR, and msImpute_MAR could not be considered in the mask-and-impute experiments (see Section 3.3).

Likewise, three remaining state-of-the-art methods could not be included in the benchmark for various reasons: (*i*) Penalized Expectation Maximization for Missing Values (PEMM) [Chen et al., 2014], which had originally been validated on a simulated dataset with 100 proteins and could not scale up to an entire real-life peptide/precursor dataset as a result of its prohibitive computational cost; (*ii*) The PIMMS methods [Webel et al., 2023] which are based on deep-learning approaches, and which consequently require large cohorts (more than 50 samples) to be fully efficient; (*iii*) ProJect [Kong et al., 2023], a general purpose omics imputation method which code has not yet been commented, structured and packaged at the time of our evaluations, so that we could not make it work on the benchmark data. (*iv*) Few elder methods with previously demonstrated poor performances (*e.g.*, zero imputation, mean imputation, accelerated failure time [Taylor et al., 2013], etc.) where not included in the benchmark, for sake of clear enough plots.

Finally, in our experiments, we did not consider any meta imputation algorithm: post-processor, like GSimp [Wei et al., 2018, Wang et al., 2022], ensemble imputation, like IMP4P [Giai Gianetto et al., 2020], chained imputation, like MICE [van Buuren and Groothuis-Oudshoorn, 2011], etc. The reasons are that their performances directly relates to the imputation algorithm they rely on, and that most of them can be extended to incorporate new input algorithms like Pirat.

3.3 Differential abundance task

On datasets with known ground-truth about differentially expressed proteins [Cox et al., 2014, Bouyssié et al., 2020, Huang et al., 2020], we compare our methods with those described in Section 3.2 using the following procedure: *(i)* Impute missing peptide (or precursor) abundance values. *(ii)* Test the all-mean equality of each peptide (or precursor) using the one-way ANOVA omnibus test available in the R package *stats* [R Core Team, 2013], and retain the resulting p-values. If a peptide cannot be imputed by a given algorithm (because of too few observed values), the test is performed nonetheless if the number of observed values allows for it, otherwise we set its p-value to one—see [R Core Team, 2013] for details about missing value restrictions. *(iii)* Display the precision-recall curves for the precision range [90%, 100%] (*i.e.*, low FDP setting) for each method, and compute the global area under the curve, (*i.e.*, on the precision range [0%, 100%]).

3.4 Mask-and-impute experiments

To precisely evaluate imputation errors, it has become customary to add pseudo-missing values with different MCAR/MNAR proportions, and to compute Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) between imputed and ground-truth values, which read as:

$$\mathrm{RMSE}(X, X^{\mathrm{imp}}) = \sqrt{\frac{1}{n} \sum_{(i,j) \in \mathcal{I}_{\mathrm{pseudo}}} (X_{i,j} - X^{\mathrm{imp}}_{i,j})^2} \ , \tag{15}$$

$$\mathrm{MAE}(X, X^{\mathrm{imp}}) = \frac{1}{n} \sum_{(i,j) \in \mathcal{I}_{\mathrm{pseudo}}} |X_{i,j} - X^{\mathrm{imp}}_{i,j}| \tag{16}$$

where $\mathcal{I}_{\mathrm{pseudo}}$ is the set of indices of pseudo-missing values and where X^{imp} is the imputed matrix. We generate pseudo MNAR values using a Probit left-censoring mechanism, *i.e.*, the probability for a value x to be missing is 1 - $\Phi(x|\nu, \tau)$, where the ν and τ are mean and standard deviation from the Gaussian cumulative distribution function Φ [Miao et al., 2016]. The overall missing value rate α and the MNAR/MCAR proportion β are controlled by applying the following procedure, adapted from Lazar et al. [2016]:

1. Set $\tau = \sigma/2$ where σ is the overall standard deviation of the dataset.
2. For each value of ν' on a linear scale between $\mu - 3\sigma$ and μ (the overall mean of the dataset), compute the expected overall rates q of missing values when applying Probit left-censoring wih parameters ν' and τ, and retain these values. In practice we compute 100 points in total.
3. Interpolate points (ν', q) obtained.

4. Compute ν associated to $\beta\alpha$ (the desired overall MNAR rate) with interpolated curve. If $\alpha\beta$ is not in the range of interpolated curve, choose larger scale for ν' at step 2.

5. Add MNAR values using Probit left-censoring with parameters ν and τ.

6. Add MCAR values with probability $\frac{\alpha(1-\beta)}{1-\alpha\beta}$.

This method ensures that:

- The overall MNAR rate is $\beta\alpha$ (by construction of interpolated curve of (ν', q)).
- The overall MV is rate is α, as

$$\begin{aligned} P(x \text{ is missing}) &= P((x \text{ is MCAR }) \cup (x \text{ is MNAR})) \\ &= P(x \text{ is MCAR}) + P(x \text{ is MNAR}) - P(x \text{ is MCAR } \cap x \text{ is MNAR}) \\ &= \frac{\alpha(1-\beta)}{1-\alpha\beta} + \beta\alpha - \frac{\beta\alpha^2(1-\beta)}{1-\alpha\beta} = \alpha\,, \end{aligned} \tag{17}$$

where x is an observed abundance value.

Of note, in the mask-and-impute experiments, all the methods are compared on the same artificially masked datasets.

4 Results

4.1 Pirat outperforms state-of-the-art on differential analysis task

We first evaluate our method on several biomarker selection tasks where significantly differentially expressed proteins are sought. We rely on three benchmark datasets for which differentially expressed proteins are known as their relative abundances are controlled, Bouyssie2020 [Bouyssié et al., 2020], Cox2014 [Cox et al., 2014], and Huang2020 [Huang et al., 2020]. They essentially involve a protein standard spiked in a complex yet constant biological background. These datasets have been produced with different MS acquisition modes (either data dependent or independent, a.k.a. DDA or DIA), at precursor- or peptide-level, and using different standard mixtures (see Supp. Mat. 3.1).

The feature selection is performed using the p-values resulting from the significance testing of the peptides' differential abundance (described in Section 3.3), after imputing with our method (Pirat) as well as with 15 different state-of-the-art or popular methods (described in Section 3.2). Although receiving operating characteristic (ROC) curves are classically used to assess feature selection (they can be found in Sup. Mat. Figure 6), we have relied on precision-recall (PR) curves (Figure 1) instead. They provide broadly similar rankings, but the PR curves display the False Discovery Proportion (FDP), which directly relates to the FDR one often controls in proteomic experiments [Burger, 2018]. More precisely, FDR being classically controlled at 1% or 5%, we present on Figure 1 partial PR curves focused on the high precision (low FDP) region to better assess selection performances in this setting. Global area under the entire PR curve (AUCPR) are also given in Table 1.

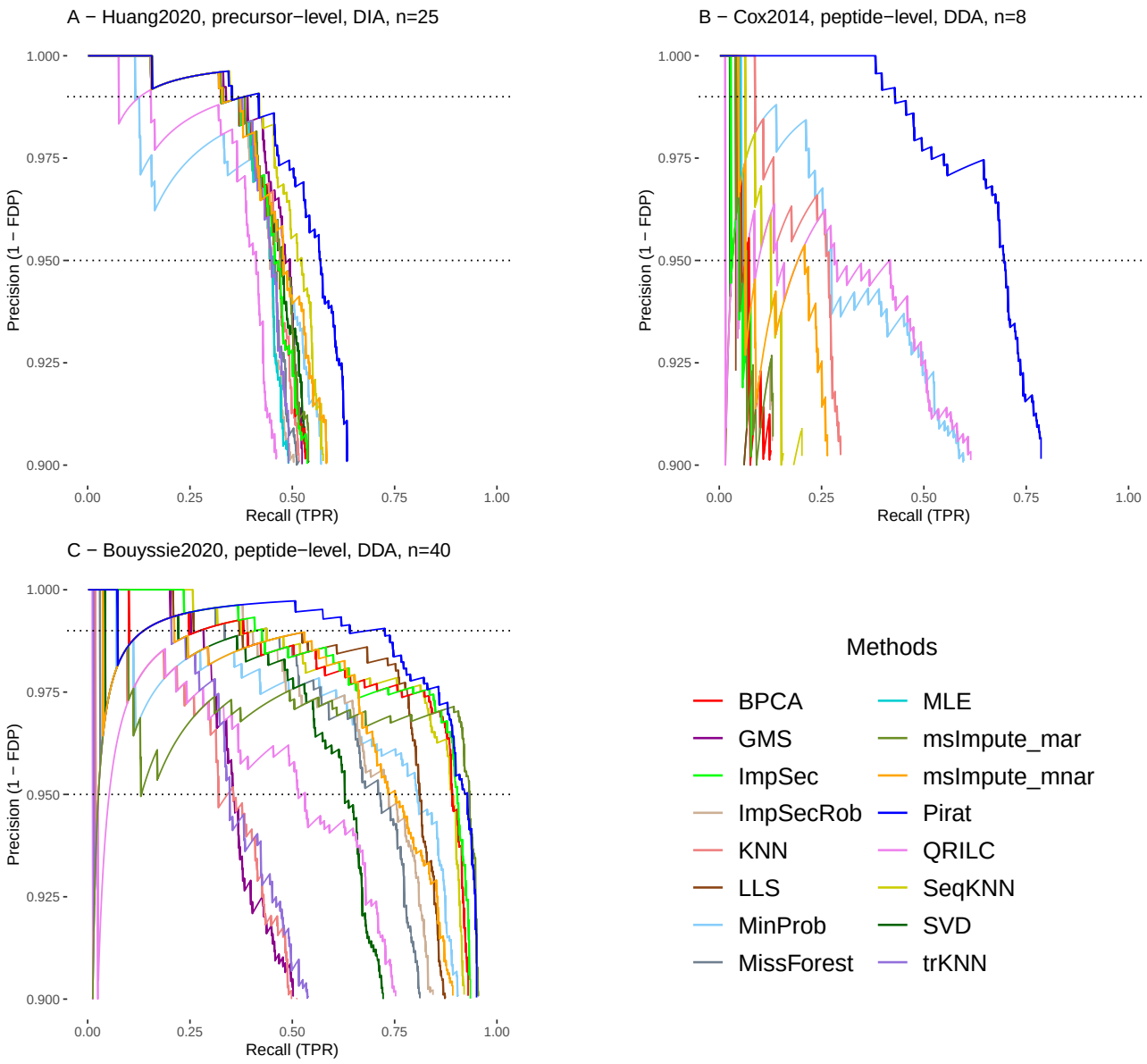


Figure 1: Differential abundance PR curves between 90% and 100% precision comparing Pirat to 15 imputation procedures on three benchmark datasets (A - Bouyssié2020, B - Cox2014, C - Huang2020), for which the name of the study, the imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are indicated in the subplot titles. The 1% and 5% FDP level are shown by the dotted lines.

On the three benchmark datasets, Pirat achieves the highest AUCPR (by a margin of 2.14% to 5.35% with respect to the second-best methods), thus indicating it best preserves the differential abundances while introducing less false positives. Precisely, Pirat provides the best PR trade-off on Huang2020 and Bouyssie2020 in the low FDP region, and almost entirely dominates the other methods on Cox2014. Beyond Pirat performances, these experiments are insightful with many respects: To begin with, several highly popular methods like MissForest, KNN or MLE have poor performances with respect to other less popular yet old methods like SeqKNN or ImpSec. Moreover, both msImpute methods (one of the most recent methods, published in April 2023) are very efficient, yet not as much as indicated in their seminal paper [Hediyeh-zadeh et al., 2023]. The reason is twofold: First, although well ranked in terms of global AUCPR, which concurs with the published results, they are outperformed by many other methods in the low FDP region (for example on Figure 1 A and C, SeqKNN exhibits a better PR tradeoff than both at low FDP). Second, they cannot impute missing values if less than four quantitative values are observed, so that the performances drop on datasets with fewer samples, like Cox2014. Conversely, Pirat clearly outperforms other methods on this dataset, which shows it can safely be applied in context of scarce samples. Most importantly, the lack of stability of other methods with respect to Pirat is noteworthy: Across the three datasets, the top scoring methods vary, first because the differences being sometimes marginal, the ranking is sensitive to random fluctuations; but also because of the changes

	AUCPR % (rank)		
Dataset	A - Huang2020	B - Cox2014	C - Bouyssié2020
BPCA	67.26 (7)	29.96	94.61 (3)
GMS	65.38 (10)	14.24	81.21
ImpSeq	67.20 (8)	31.20	95.08 (2)
ImpSeqRob	64.11	29.74	90.47 (9)
KNN	59.33	51.94 (6)	79.94
LLS	SegFault	31.48 (10)	92.19 (8)
MinProb	72.56 (3)	81.08 (2)	93.91 (6)
MissForest	63.87	3.70	88.91
MLE	62.77	34.46 (8)	23.58
msImpute_mar	71.78 (4)	52.31 (5)	94.22 (4)
msImpute_mnar	73.88 (2)	68.92 (4)	94.14 (5)
Pirat	**77.02 (1)**	**86.43 (1)**	**97.22 (1)**
QRILC	68.08 (5)	80.76 (3)	89.45 (10)
SeqKNN	66.60 (9)	34.91 (7)	93.81 (7)
SVD	67.32 (6)	31.83 (9)	89.03
trKNN	64.15	14.23	80.71

Table 1: Global area under the precision-recall curves (and ranking into brackets) comparing Pirat with 15 imputation algorithms on a differential analysis task using three benchmark datasets (Bouyssié2020, Cox2014, Huang2020).

in the MCAR/MNAR proportion. On Bouyssie2020, the MAR/MCAR oriented methods perform best after Pirat (ImpSeq, BPCA, and msImpute_mar), whereas on Cox2014 and Huang2020, Pirat is followed by MNAR oriented methods (MinProb, msImpute_mnar, QRILC). Overall, only Pirat remains in the top 3 for all datasets, and it does so with the first rank.

4.2 Pirat is robust with respect to its underpinning assumptions

Although Pirat displays the highest performances in an end-to-end task like biomarker discovery, understanding the extent to which the performance increments roots in the MS left-censoring or in the sibling peptide (peptides belonging to the same PG) correlations is insightful. To do so, we rely on mask-and-impute experiments (*i.e.*, missing values are artificially added in real datasets, so that it is possible to measure the difference between the masked values and their imputed counterparts). To provide a baseline to the evaluations, we compare to the best-ranked methods according to the previous experiments. However, we have not been able to include msImpute_mnar, msImpute_mar, GMS, and trKNN irrespective of their performances, because of restriction on the number of observed values (see Section 3.2). Therefore, the subsequent evaluations compare Pirat with the best performing methods which can handle peptides with at least two observed values: MinProb, QRILC (MNAR scenario) and ImpSec, BPCA, SeqKNN (MAR/MCAR scenarios). Finally, a degenerated version of Pirat is also considered, where peptides are processed individually (*i.e.*, in PGs or size 1), so that within-PG correlations cannot be leveraged.

In this experiment, we rely on datasets from two other LC-MS/MS proteomic analyses, detailed in Section 3.1. The first one, Capizzi2022 [Capizzi et al., 2022], is composed of 10 samples processed at the peptide level (16% of MVs). The second and third ones, Vilallongue2022 SCN and SC [Vilallongue et al., 2022], contain 8 samples processed at the precursor level (respectively 14% and 16% of MVs). According to our diagnosis tool (see Figure 2 second row), the sibling peptides tend to be strongly and positively correlated in Capizzi2022, as expected from a normal proteomic experiment. Oppositely, in Vilallongue2022 SCN, sibling peptides depict an odd correlation profile, and even more so in Vilallongue2022 SC, where the assumptions of bottom-up proteomics are hardly verified. Although questioning from an analytical viewpoint, these datasets are insightful, as they enable to evaluate PG's benefits when siblings correlations are poor. For all datasets, we only consider the subset of peptides or precursors with no missing value, to which we introduce artificial MVs at a rate equal to that of the original dataset. Finally, we filter out peptides or precursors with one or zero remaining observed values. To evaluate the impact of the censoring model, which aims at improving MNAR imputation, these artificial MVs are a mix of MCARs and MNARs, where the former ones are generated according to a Probit mechanism [Miao et al., 2016], as described in Section 3.4. Note that this model used for masking is different from Pirat's missingness model, which permits evaluating it in a misspecified setting. Then, we impute the missing values using the algorithms to benchmark and compute their associated Root Mean Square Error (RMSE) and Mean Absolute Error (MAE, given in Supp. Mat. A.5). We repeat this operation for different MNAR proportions (0%, 25%, 50%, 75%, and 100%) and seeds, leading to the curves of Figure 2.

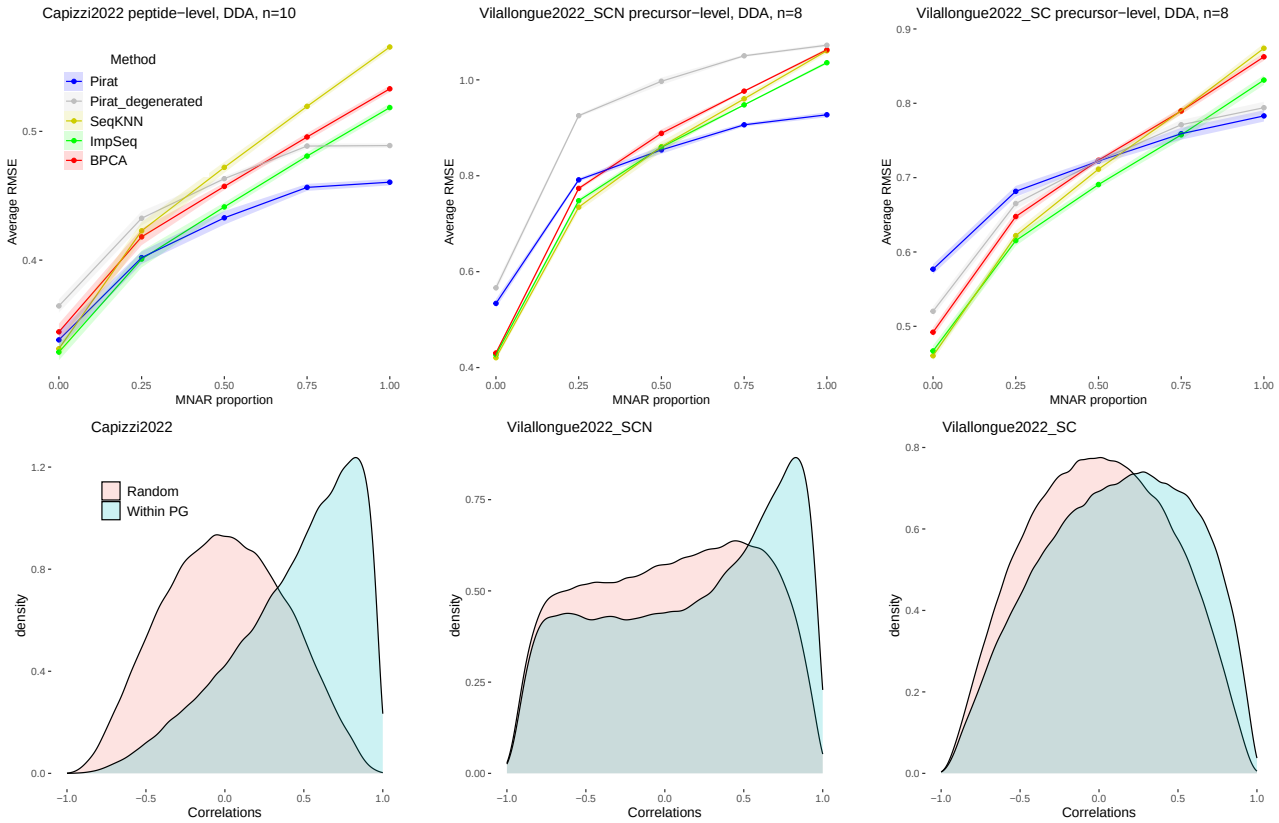


Figure 2: Average RMSE (top) of best imputation methods (according to the previous results) in function of the proportion of MNAR values on Capizzi2022 (left), Vilallongue2022 SCN (center), and Vilallongue2022 SC (right). The imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are indicated in the subplot titles. The errors averaged over 5 different seeds and margins correspond to standard deviations. On the bottom, empirical densities of correlations between peptides chosen randomly and between sibling peptides for each dataset respectively.

We do not display QRILC and MinProb results in Figure 2 as their MAE and RMSE is 3 to 4 times higher than that of other methods (full plots available in Sup. Mat. Figure 8). For those methods, the difference of performances between mask-and-impute and differential analysis is striking although easily explainable: these MNAR-devoted methods are highly biased towards low abundances and work in a univariate setting (*i.e.*, apart from the knowledge of an ill-defined instrumental limit, no information is used), thus resulting in poor quantitative predictions. Yet, in cases where this low abundance is biologically relevant, erroneous predictions may not impede the accuracy of differential analysis: For instance, when a protein is not expressed in a phenotype, imputing its peptides with extremely low values may cause no significant harm.

All the other reference methods tested (PBCA, ImpSeq, and SeqKNN) have similar trends in terms of RMSE and MAE: As MCAR-devoted methods, the associated RMSE and MAE increase with the MNAR proportion. Although these three methods are broadly equally performing, ImpSec seems slightly but consistently more accurate, regardless of the MNAR ratio. As for Pirat, as expected, the performances vary according to the MNAR ratio and the quality of within-PG correlations, which validate the assumption at the root of Pirat algorithm. In more details, this variety of scenarios pinpoints the weaknesses of Pirat: in absence of MNAR values (which is not realistic in LC-MS/MS analyses), the performances are not as good because the instrument censoring model is useless. Likewise, in case of too low within-PG correlations, the performances decrease. However, when one out of the two assumption is fulfilled (significant MNAR ratio and poor within-PG correlations, or conversely, high within-PG correlations and large MCAR ratio) the performances are equivalent to the best state-of-the-art methods, both in terms of MAE and RMSE. More interestingly, as soon as both assumptions are fulfilled (which should be expected from most proteomics experiments), Pirat's performances are unmatched, which explains why it has outperformed all other methods in the differential analysis task (see Section 3.3). Finally, we observe that Pirat outperforms in most cases its degenerated version. This is particularly showcased in MNAR dominant settings, which highlights the benefits of relying on PGs. On the opposite, it unveils the risk of lower-quality imputation for weakly-covered proteins.

4.3 Improving peptide imputation for proteins with low coverage

We now evaluate various options to best process the peptides of weakly-covered proteins. Precisely, we focus on singleton PGs, which are the most difficult situations for Pirat. As borderline cases, there is no unique satisfactory way to process them. The most straightforward and most statistically-grounded approach would be to adopt a quantitative version of the two-peptide rule often used at peptide identification [Munteanu et al., 2018], which would lead to filter out singleton PGs with missing values (see Supp. Mat. A.7). In case doing so is not acceptable for experimental reasons, we propose two alternatives: The first one (referred to as Pirat-S, see Section 2.9.1) is to leverage sample-wise correlations for imputation, with a similar logic as other algorithms (*e.g.*, like ImpSeq or BPCA). The second one (Pirat-T, see Section 2.9.1) is to rely on complementary transcriptomic analyses, opening the path to a multi-omic view on imputation.

To evaluate them, we rely on two datasets involving both label-free proteomics and transcriptomics: Ropers2021 [Ropers et al., 2021] and Habowski2020 [Habowski et al., 2020]. Ropers et al. [2021] studied the effect of a synthetic growth switch on *E. coli* by performing paired proteomic and transcriptomic analyses of a wild-type and mutated strain at respectively 2 and 4 timepoints, thus resulting in 6 conditions (3 replicates each). In the mutated strain, protein and mRNA expressions are expected to decrease over time at different rate due to the differences of half-life of mRNA and proteins. As such, the samples have similar compositions leading to a relatively low MV rate at precursor-level (about 15%,) and we expect the transcriptomic-proteomic correlations to be exploitable thanks to the paired design.

In contrast, Habowski2020 results from unpaired transcriptomic and proteomic analyses of six types of mouse colon cells (3 replicates each for the proteomic experiments and 2 to 5 for the transcriptomic ones). Each condition being a different cell type, the sample composition in proteins varies considerably between the different conditions. This results in an important amount of precursor-level MVs (which reaches 50%) with values missing on an entire condition, a.k.a. MEC [Wieczorek et al., 2019]. Note that, compared to Ropers2021, this second dataset is crucial to (*i*) measure the possible negative impact of uncorrelated transcriptomic data in a multi-omic imputation approach; (*ii*) assess the importance of within-PG correlations to correctly impute poorly observed peptides; (*iii*) to evaluate if the performances are affected when samples are not perfectly correlated (leading to MEC-type MNARs). See Sup. Mat. 3.1 for more details on both datasets.

To evaluate the performance of these approaches, we rely on a mask-and-impute experiment similar to the previous one, with the additional goal to perturbate the true distribution of MVs the least amount possible. For this reason, we use the entire set of precursors (instead of using precursors with no MVs) and we add only a small amount of MVs (1% of the total number of true MVs in each dataset). This results in a few values that can be used to assess the performance of the approaches. We thus consider only the two extreme scenarios: 0% and 100% MNARs (respectively named MCAR and MNAR scenarios in the following experiments) and repeat this process ten times to assess the overall distribution of errors in each scenario.

4.3.1 Pirat-S

Figure 3 shows, among others, the absolute errors of Pirat and Pirat-S. Clearly, Pirat-S improves upon Pirat in the MCAR setting for both datasets. This is not surprising: Pirat-S is conceptually close to ImpSeq, which has demonstrated excellent performances on MCARs (see Section 4.2). The improvement is more nuanced in the MNAR setting. Specifically, it seems to deteriorate the performances on Habowski2020. As pseudo-MNAR MVs are mostly carried by empty peptides (MEC patterns), too few values are left to derive informative correlations, notably when they do not originate from the same sample (see Sup. Mat. Figure 10, where the histogram E has a clearly different trend than histograms F, G and H). Although Pirat-S seems to be a safe choice in all other cases, medium-only imputation performances on MNARs with MEC (or similar) patterns reveal a limit of this approach.

4.3.2 Pirat-T

Intuitively, the performance increment between Pirat and Pirat-T is expected to depend on the correlations between peptides and transcripts abundances. This is verified on Figure 3: the distribution error is shifted downwards on Ropers2021, where proteomic and transcriptomic samples are paired. On the contrary, the differences seem marginal on Habowski2020, where samples are not paired, thus making the transcriptomic patterns hardly exploitable on the proteomic data. Importantly, although transcriptomics is not helpful in this case, it does not deteriorate the performances either. This indicates that even in case a researcher wonders about the transcriptomic data quality or about their correlations with proteomics, it is not riskier to incorporate them for downstream imputation: Pirat-T robustly integrates transcriptomic data for imputation, without overinterpreting mRNA variations.

When comparing Pirat-S and Pirat-T, it appears the latter do not outperform the former, indicating that in general, sample-wise correlations can be more robustly leveraged than transcriptomic-proteomic ones (see

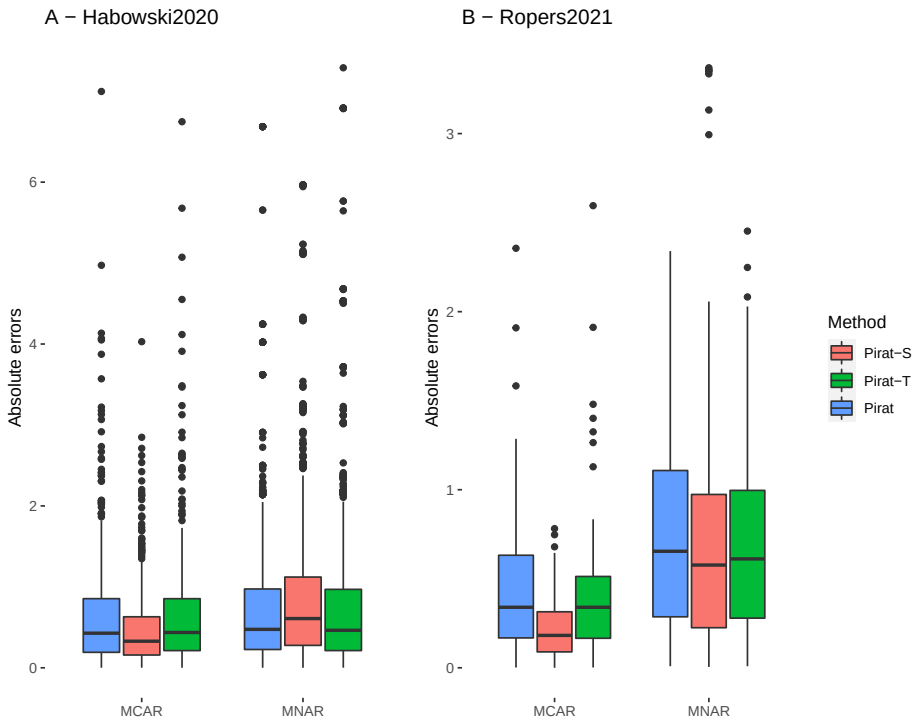


Figure 3: Boxplot of absolute errors for singleton PGs in A - Habowski202 and B - Ropers2021 datasets of Pirat, Pirat-S and Pirat-T.

Fortelny et al. [2017] for a discussion related to this topic). Although disappointing from a multi-omic integration perspective, this conclusion suffers one noteworthy exception: Habowski2020 on the MNAR setting; and interestingly, this is observed despite the lack of paired design. This suggests that in case of analysing proteomic samples from different tissues (with MEC-like MV values, which cannot be imputed accurately by leveraging sample or sibling peptide correlations) imputation can advantageously rely on transcriptomic assays. As those MEC values are essentially MNARs, it confirms the intuition shared in the proteomic community that complementary omic studies can be effective sources of information to explore more safely the proteome beyond the quantitation limit of mass spectrometry.

5 Discussion

Overall, Pirat outperforms previously published imputation methods across various experimental settings. On datasets endowed with differential abundance ground truth labels but with different missingness patterns and replicate numbers, Pirat consistently achieves the best PR trade-off, whereas other competitive methods display less stable and less accurate performances. Likewise, on mask-and-impute experiments, Pirat exhibits lower imputation errors for all scenarios with a majority of MNARs or with sufficiently good sibling peptide correlations. As for the other less realistic scenarios, the performances are more variable, yet close to the best ones in the worst cases. Indeed, while MCAR-devoted methods has long dominated the field of proteomics imputation, ignoring the physical limits of the instrument can only lead to biased results. Likewise, even though "discordant" peptides among siblings are likely to be observed as a result of unknown post-translational modifications or of poorly quantified peptides [Dermit and Meyer, 2021], they should be too scarce to affect the correlation distribution. Therefore, too low correlations should raise the experimenter's awareness about the biochemical consistency of the assay. With this extent, the diagnosis tool accompanying Pirat is insightful. The above case suffers one exception, peptides without siblings (*i.e.*, in singleton PGs), for which the difficulty to impute using Pirat does not relate to the experiment quality. To process them, we propose three alternatives to the original approach: (*i*) the quantitative two-peptide rule, *i.e.*, discarding proteins with a single not fully quantified peptide. (*ii*) If different tissues are compared, it can be insightful to complement the assay using paired transcriptomics, as then Pirat-T can efficiently enhance the imputation of MNARs on singleton PGs. (*iii*) If transcriptomics is not available or not relevant for the proteomic study and if similar samples are analysed, sample-wise correlations can improve singleton peptide imputation using Pirat-S. With this regard, we draw attention to the marginal differences of global performances between the various extention of Pirat: Once averaged on an entire dataset, the imputation error on singleton PGs alone does not weight as much as the experimental design to drive the imputation strategy. We therefore encourage Pirat's users to question their

need with respect to proteins with low coverage before opting which strategy to use. Lastly, our experiments have reported Pirat's excellent performances on peptide- and precursor-level datasets, thus ensuring its safe use in both cases. These results unfortunately do not allow us to formulate an educated opinion about which level is most appropriate to impute. Closing this debate would require testing the two approaches with a metric authorizing their comparison.

6 Conclusions

Considering its performances, we believe future research in proteomics imputation would advantageously leverage Pirat's paradigms. Let us recall them: (*i*) the estimation of a global missingness mechanism inferred from the data; (*ii*) an explicit modeling of the biochemical dependencies that are known to result from the analytical pipeline; (*iii*) the possibility to include other omic measurements that are expected to correlate to the proteomic ones. Pushing further the developments in either of these three trends may require a different mathematical backbone than that of Pirat (*i.e.*, not necessarily involving the penalized likelihood model of Equation 6), however, multiple paths are promising: regarding the missingness mechanism, a natural parameter estimator reads as the maximum likelihood estimate (MLE) over all the PGs. However, it requires replacing the optimization procedure described in Section 2.7 by a joint MLE, which memory and computation cost would be prohibitive. Fortunately, as Pirat processes each PG independently, it is highly parallelizable if acceleration is needed. Also, other missingness patterns like Probit or Logit [Li and Smyth, 2023] can fit the data, but before investigating them, some guarantees are necessary about their log-likelihood (or any lower bound, see Section 2.7) being tractable and differentiable. Besides, the self-masked assumption on the missingness mechanism only holds if we neglect the so-called "dynamic range" of mass spectrometers [Zubarev, 2013]. In practice, the variation of lower detection bound are not purely stochastic: they also result from which peptides are currently measured, notably the most abundant ones (as if highly abundant peptides could mask lower abundant ones). Thus, accounting for the intensity of co-eluting peptides in the missingness mechanism would be probably more accurate. A natural solution would be to add an extra term including relevant peptide abundances in Equation 5. However, this would require having access to additional data that are classically not reported, and extracting those pertaining to the dynamic range effect. From a biological standpoint, databases of known protein-protein interactions (PPI) are a source of supplementary correlation patterns, and incorporating them can be instrumental, notably for singleton PGs (as to find them pseudo-siblings capable of guiding the imputation). Likewise, known biologically or analytically relevant interactions can also be encoded in the prior of the covariance matrix Σ. For example, the correlations between protein-specific sibling peptides are expected to be larger than those involving either shared peptides, transcripts or PPI, so that it would make sense to adapt accordingly the scaling matrix prior described in Section 2.6. Finally, Pirat proposes to incorporate transcriptomic data, but other high-throughput technologies provide access to gene-wise or PG-wise quantitative information, which integration into Pirat follows the same logic (broadly speaking, one just appends the quantitative vectors of the various omics into a single matrix). Therefore, Pirat naturally opens the path to more elaborated integrative multi-omic methods to further tackle the challenge of proteomic data imputation.

Competing interests

No competing interest is declared.

Author contributions statement

Funding acquisition (TB), research design and supervision (NV, TB), methodological development and prototyping (LE), data preparation (LE, LF, TB), experimental comparisons (LE), result analysis (LE, NV, TB), code and packaging (LE, SW), manuscript drafting (LE, NV, TB), manuscript assessment (LE, LF, SW, NV, TB).

Acknowledgments

This work was supported by grants from the French National Research Agency: ProFI project (ANR-10-INBS-08), GRAL CBH project (ANR-17-EURE-0003), SECRET project (ANR-22-CE45-0026), DEAP project (ANR-15-IDEX-02), and MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

Availability of data and materials

The data and the methods are available on a publicly accessible Github repositories:

- `https://github.com/prostarproteomics/Pirat`
- `https://github.com/TrEE-TIMC/Pirat_experiments`

References

S. Anders, P. T. Pyl, and W. Huber. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, jan 2015. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU638. URL `https://academic.oup.com/bioinformatics/article/31/2/166/2366196`.

R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, jun 2018. ISSN 1744-4292. doi:10.15252/msb.20178124. URL `https://www.embopress.org/doi/full/10.15252/msb.20178124https://www.embopress.org/doi/abs/10.15252/msb.20178124`.

M. P. Barzine, K. Freivalds, J. C. Wright, M. Opmanis, D. Rituma, F. Z. Ghavidel, A. F. Jarnuczak, E. Celms, K. Čerāns, I. Jonassen, L. Lace, J. Antonio Vizcaíno, J. S. Choudhary, A. Brazma, and J. Viksna. Using Deep Learning to Extrapolate Protein Expression Measurements. *Proteomics*, 20(21-22), 2020. ISSN 16159861. doi:10.1002/pmic.202000009.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU170. URL `https://academic.oup.com/bioinformatics/article/30/15/2114/2390096`.

D. Bouyssié, A. M. Hesse, E. Mouton-Barbosa, M. Rompais, C. MacRon, C. Carapito, A. Gonzalez De Peredo, Y. Couté, V. Dupierris, A. Burel, J. P. Menetrey, A. Kalaitzakis, J. Poisat, A. Romdhani, O. Burlet-Schiltz, S. Cianférani, J. Garin, and C. Bruley. Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics*, 36(10):3148–3155, may 2020. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTAA118. URL `https://academic.oup.com/bioinformatics/article/36/10/3148/5756203`.

K. V. Branden and S. Verboven. Robust data imputation. *Computational Biology and Chemistry*, 33(1):7–13, feb 2009. ISSN 1476-9271. doi:10.1016/J.COMPBIOLCHEM.2008.07.019.

T. Burger. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *Journal of Proteome Research*, 17(1):12–22, jan 2018. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.7B00170. URL `https://pubs.acs.org/doi/full/10.1021/acs.jproteome.7b00170`.

M. Capizzi, R. Carpentier, E. Denarier, A. Adrait, R. Kassem, M. Mapelli, Y. Couté, and S. Humbert. Developmental defects in Huntington's disease show that axonal growth and microtubule reorganization require NUMA1. *Neuron*, 110(1):36–50.e5, jan 2022. ISSN 0896-6273. doi:10.1016/J.NEURON.2021.10.033.

L. S. Chen, R. L. Prentice, and P. Wang. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*, 70(2):312–322, jun 2014. ISSN 1541-0420. doi:10.1111/BIOM.12149. URL `https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12149https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12149https://onlinelibrary.wiley.com/doi/10.1111/biom.12149`.

M. Chion and A. Leroy. A bayesian framework for multivariate differential analysis accounting for missing data, 2023.

M. Chion, C. Carapito, and F. Bertrand. Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics. *PLOS Computational Biology*, 18(8): e1010420, aug 2022. ISSN 1553-7358. doi:10.1371/journal.pcbi.1010420. URL `https://dx.plos.org/10.1371/journal.pcbi.1010420`.

J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics : MCP*, 13(9):2513, sep 2014. ISSN 15359484.

doi:10.1074/MCP.M113.031591. URL /pmc/articles/PMC4159666//pmc/articles/PMC4159666/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159666/.

M. Dermit and J. G. Meyer. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *Journal of Proteome Research*, 2021. ISSN 15353907. doi:10.1021/acs.jproteome.0c00602.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTS635. URL https://academic.oup.com/bioinformatics/article/29/1/15/272537.

B. Efron and C. Morris. Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, 67(337):130–139, mar 1972. ISSN 0162-1459. doi:10.1080/01621459.1972.10481215. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481215http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481215.

J. E. Flores, D. M. Claborne, Z. D. Weller, B. J. M. Webb-Robertson, K. M. Waters, and L. M. Bramer. Missing data in multi-omics integration: Recent advances through artificial intelligence. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 26248212. doi:10.3389/frai.2023.1098308.

N. Fortelny, C. M. Overall, P. Pavlidis, and G. V. C. Freue. Can we predict protein from mRNA levels? *Nature 2017 547:7664*, 547(7664):E19–E20, jul 2017. ISSN 1476-4687. doi:10.1038/nature22293. URL https://www.nature.com/articles/nature23293.

M. L. Gardner and M. A. Freitas. Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics. *International Journal of Molecular Sciences*, 22(17):9650, sep 2021. ISSN 1422-0067. doi:10.3390/ijms22179650. URL https://www.mdpi.com/1422-0067/22/17/9650.

Q. Giai Gianetto, S. Wieczorek, Y. Couté, and T. Burger. A peptide-level multiple imputation strategy accounting for the different natures of missing values in proteomics data. *bioRxiv*, page 2020.05.29.122770, may 2020. doi:10.1101/2020.05.29.122770. URL https://www.biorxiv.org/content/10.1101/2020.05.29.122770v1https://www.biorxiv.org/content/10.1101/2020.05.29.122770v1.abstract.

L. J. E. Goeminne, A. Sticker, L. Martens, K. Gevaert, and L. Clement. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Analytical Chemistry*, 92(9):6278–6287, may 2020. ISSN 0003-2700. doi:10.1021/acs.analchem.9b04375. URL https://pubs.acs.org/doi/10.1021/acs.analchem.9b04375.

S. Gupta, C. N. Schlaffner, S. Ahmed, H. Steen, B. Y. Renard, and K. Baum. Imputing protein abundance by modeling molecular relationships using gnns. https://www.iscb.org/ismbeccb2023-programme/posters, 2023. Accessed: 2023-10-05.

A. N. Habowski, J. L. Flesher, J. M. Bates, C. F. Tsai, K. Martin, R. Zhao, A. K. Ganesan, R. A. Edwards, T. Shi, H. S. Wiley, Y. Shi, K. J. Hertel, and M. L. Waterman. Transcriptomic and proteomic signatures of stemness and differentiation in the colon crypt. *Communications Biology 2020 3:1*, 3(1):1–17, aug 2020. ISSN 2399-3642. doi:10.1038/s42003-020-01181-z. URL https://www.nature.com/articles/s42003-020-01181-z.

S. Hediyeh-zadeh, A. I. Webb, and M. J. Davis. MsImpute: Estimation of missing peptide intensity data in label-free quantitative mass spectrometry. *Molecular & Cellular Proteomics*, page 100558, apr 2023. ISSN 1535-9476. doi:10.1016/J.MCPRO.2023.100558. URL https://linkinghub.elsevier.com/retrieve/pii/S1535947623000683.

T. Huang, R. Bruderer, J. Muntel, Y. Xuan, O. Vitek, and L. Reiter. Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition. *Molecular & Cellular Proteomics*, 19(2):421–430, feb 2020. ISSN 1535-9476. doi:10.1074/MCP.RA119.001705.

L. Jin, Y. Bi, C. Hu, J. Qu, S. Shen, X. Wang, and Y. Tian. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11(1):1760, dec 2021. ISSN 20452322. doi:10.1038/s41598-021-81279-4. URL https://doi.org/10.1038/s41598-021-81279-4.

Y. V. Karpievitch, A. R. Dabney, and R. D. Smith. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics 2012 13:16*, 13(16):1–9, nov 2012. ISSN 1471-2105. doi:10.1186/1471-2105-13-S16-S5. URL https://link.springer.com/articles/10.1186/1471-2105-13-S16-S5https://link.springer.com/article/10.1186/1471-2105-13-S16-S5.

H. Kim, G. H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, jan 2005. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTH499. URL https://dx.doi.org/10.1093/bioinformatics/bth499.

K. Y. Kim, B. J. Kim, and G. S. Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, 5(1):1–9, oct 2004. ISSN 14712105. doi:10.1186/1471-2105-5-160. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-160.

W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh. Dealing with missing values in proteomics data. *PROTEOMICS*, page 2200092, nov 2022. ISSN 1615-9853. doi:10.1002/PMIC.202200092. URL https://onlinelibrary.wiley.com/doi/10.1002/pmic.202200092.

W. Kong, B. J. H. Wong, H. W. H. Hui, K. P. Lim, Y. Wang, L. Wong, and W. W. B. Goh. ProJect: a powerful mixed-model missing value imputation method. *Briefings in Bioinformatics*, 24(4), jul 2023. ISSN 14774054. doi:10.1093/BIB/BBAD233. URL https://dx.doi.org/10.1093/bib/bbad233.

C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. 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. ISSN 15353907. doi:10.1021/acs.jproteome.5b00981.

E. Leppäaho, S. Kaski, and M. E. Khan. GFA: Exploratory Analysis of Multiple Data Sources with Group Factor Analysis Muhammad Ammad-ud-din. *Journal of Machine Learning Research*, 18:1–5, 2017. doi:10.5555/3122009.3122048. URL http://jmlr.org/papers/v18/16-509.html.

M. Li and G. K. Smyth. Neither random nor censored: estimating intensity-dependent probabilities for missing values in label-free proteomics. *Bioinformatics*, 39(5), may 2023. ISSN 1367-4811. doi:10.1093/bioinformatics/btad200. URL https://dx.doi.org/10.1093/bioinformatics/btad200https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad200/7126416.

Q. Li, K. Fisher, W. Meng, B. Fang, E. Welsh, E. B. Haura, J. M. Koomen, S. A. Eschrich, B. L. Fridley, and Y. A. Chen. GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. *Bioinformatics*, 36(1):257–263, jan 2020. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTZ488. URL https://dx.doi.org/10.1093/bioinformatics/btz488.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. doi:10.1002/9781119482260.

D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, aug 1989. ISSN 00255610. doi:10.1007/BF01589116/METRICS. URL https://link.springer.com/article/10.1007/BF01589116.

M. Liu and A. Dongre. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Briefings in Bioinformatics*, 22(3):1–17, 2021. ISSN 1467-5463. doi:10.1093/bib/bbaa112.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, dec 2014. ISSN 1474760X. doi:10.1186/S13059-014-0550-8. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

W. Ma, S. Kim, S. Chowdhury, Z. Li, M. Yang, S. Yoo, F. Petralia, J. Jacobsen, J. J. Li, X. Ge, K. Li, T. Yu, N. Edwards, S. Payne, P. Boutros, H. Rodriguez, G. Stolovitzky, J. Zhu, J. Kang, D. Fenyo, J. Saez-Rodriguez, and P. Wang. DreamAI: algorithm for the imputation of proteomics data. *bioRxiv*, page 2020.07.21.214205, jul 2020. doi:10.1101/2020.07.21.214205. URL https://doi.org/10.1101/2020.07.21.214205.

C. Meng, A. Basunia, B. Peters, A. M. Gholami, B. Kuster, and A. C. Culhane. MOGSA: Integrative single sample gene-set analysis of multiple omics data. *Molecular and Cellular Proteomics*, 18(8): S153–S168, aug 2019. ISSN 15359484. doi:10.1074/mcp.TIR118.001251. URL http://www.mcponline.org/article/S1535947620327687/fulltexthttp://www.mcponline.org/article/S1535947620327687/abstracthttps://www.mcponline.org/article/S1535-9476(20)32768-7/abstract.

W. Miao, P. Ding, and Z. Geng. Identifiability of Normal and Normal Mixture Models with Nonignorable Missing Data. *Journal of the American Statistical Association*, 111(516):1673–1683, oct 2016. ISSN 1537274X. doi:10.1080/01621459.2015.1105808. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2015.1105808.

M. Munteanu, R. Pais, V. Peta, O. Deckmyn, J. Moussalli, Y. Ngo, M. Rudler, P. Lebray, F. Charlotte, V. Thibault, O. Lucidarme, A. Ngo, F. Imbert-Bismut, C. Housset, D. Thabut, V. Ratziu, and T. Poynard. Long-term prognostic value of the FibroTest in patients with non-alcoholic fatty liver disease, compared to chronic hepatitis C, B, and alcoholic liver disease. *Alimentary Pharmacology & Therapeutics*, 48(10):1117–1127, nov 2018. ISSN 1365-2036. doi:10.1111/APT.14990. URL https://onlinelibrary.wiley.com/doi/full/10.1111/apt.14990https://onlinelibrary.wiley.com/doi/abs/10.1111/apt.14990https://onlinelibrary.wiley.com/doi/10.1111/apt.14990.

S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. I. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, nov 2003. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTG287. URL https://dx.doi.org/10.1093/bioinformatics/btg287.

J. J. O'brien, H. P. Gunawardena, J. A. Paulo, X. Chen, J. G. Ibrahim, S. P. Gygi, and B. F. Qaqish. The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, 12(4):2075, dec 2018. ISSN 19417330. doi:10.1214/18-AOAS1144. URL /pmc/articles/PMC6249692//pmc/articles/PMC6249692/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249692/.

J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing 1996*, 6(3):289–296, 1996. ISSN 1573-1375. doi:10.1007/BF00140873. URL https://link.springer.com/article/10.1007/BF00140873.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

C. P. Robert and G. Casella. *Monte Carlo Integration*, pages 71–138. Springer New York, New York, NY, 1999. ISBN 978-1-4757-3071-5. doi:10.1007/978-1-4757-3071-5_3. URL https://doi.org/10.1007/978-1-4757-3071-5_3.

F. Rohart, B. Gautier, A. Singh, and K.-A. L. Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, nov 2017. ISSN 1553-7358. doi:10.1371/JOURNAL.PCBI.1005752. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752.

D. Ropers, Y. Couté, L. Faure, S. Ferré, D. Labourdette, A. Shabani, L. Trouilh, P. Vasseur, G. Corre, M. Ferro, M. A. Teste, J. Geiselmann, and H. de Jong. Multiomics Study of Bacterial Growth Arrest in a Synthetic Biology Application. *ACS Synthetic Biology*, 10(11):2910–2926, nov 2021. ISSN 21615063. doi:10.1021/ACSSYNBIO.1C00115. URL https://pubs.acs.org/doi/full/10.1021/acssynbio.1c00115.

S. Y. Ryu, W. J. Qian, D. G. Camp, R. D. Smith, R. G. Tompkins, R. W. Davis, and W. Xiao. Detecting differential protein expression in large-scale population proteomics. *Bioinformatics*, 30(19):2741–2746, oct 2014. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU341. URL https://dx.doi.org/10.1093/bioinformatics/btu341.

J. S. Shah, S. N. Rai, A. P. DeFilippis, B. G. Hill, A. Bhatnagar, and G. N. Brock. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics*, 18(1):1–13, feb 2017. ISSN 14712105. doi:10.1186/S12859-017-1547-6/TABLES/4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1547-6.

M. Shen, Y. T. Chang, C. T. Wu, S. J. Parker, G. Saylor, Y. Wang, G. Yu, J. E. Van Eyk, R. Clarke, D. M. Herrington, and Y. Wang. Comparative assessment and novel strategy on methods for imputing proteomics data. *Scientific Reports 2022 12:1*, 12(1):1–11, jan 2022. ISSN 2045-2322. doi:10.1038/s41598-022-04938-0. URL https://www.nature.com/articles/s41598-022-04938-0.

H.-J. M. Shi and D. Mudigere. Pytorch - lbfgs. https://github.com/hjmshi/PyTorch-LBFGS, 2020.

A. Sportisse, C. Boyer, and J. Josse. Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020. doi:10.5555/3495724.3496317. URL https://github.com/AudeSportisse/PPCA_MNAR.

D. J. Stekhoven and P. Bühlmann. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, jan 2012. ISSN 13674803. doi:10.1093/BIOINFORMATICS/BTR597.

S. L. Taylor, G. S. Leiserowitz, and K. Kim. Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry 'Omic Studies. *Statistical applications in genetics and molecular biology*, 12(6):703, dec 2013. doi:10.1515/SAGMB-2013-0021. URL /pmc/articles/PMC3905689//pmc/articles/PMC3905689/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3905689/.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, jun 2001. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/17.6.520. URL https://dx.doi.org/10.1093/bioinformatics/17.6.520.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi:10.18637/jss.v045.i03.

S. Verboven, K. V. Branden, and P. Goos. Sequential imputation for missing values. *Computational Biology and Chemistry*, 31(5-6):320–327, oct 2007. ISSN 1476-9271. doi:10.1016/J.COMPBIOLCHEM.2007.07.001.

V. Vidova and Z. Spacil. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Analytica Chimica Acta*, 964:7–23, apr 2017. ISSN 0003-2670. doi:10.1016/J.ACA.2017.01.059.

N. Vilallongue, J. Schaeffer, A. M. Hesse, C. Delpech, B. Blot, A. Paccard, E. Plissonnier, B. Excoffier, Y. Couté, S. Belin, and H. Nawabi. Guidance landscapes unveiled by quantitative proteomics to control reinnervation in adult visual system. *Nature Communications*, 13(1):1–20, oct 2022. ISSN 20411723. doi:10.1038/s41467-022-33799-4. URL https://www.nature.com/articles/s41467-022-33799-4.

J. Wang, X. Gong, M. Hu, and L. Zhao. Improved GSimp: A Flexible Missing Value Imputation Method to Support Regulatory Bioequivalence Assessment. *Annals of Biomedical Engineering 2022*, pages 1–11, sep 2022. ISSN 1573-9686. doi:10.1007/S10439-022-03070-4. URL https://link.springer.com/article/10.1007/s10439-022-03070-4.

S. Wang, W. Li, L. Hu, J. Cheng, H. Yang, and Y. Liu. NAguideR: Performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Research*, 48(14):e83, aug 2020. ISSN 13624962. doi:10.1093/nar/gkaa498. URL /pmc/articles/PMC7641313//pmc/articles/PMC7641313/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7641313/.

B. J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, and K. M. Waters. 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. ISSN 15353907. doi:10.1021/pr501138h.

H. Webel, L. Niu, A. B. Nielsen, M. Locard-Paulet, M. Mann, L. J. Jensen, and S. Rasmussen. Imputation of label-free quantitative mass spectrometry-based proteomics data using self supervised deep learning. *bioRxiv*, page 2023.01.12.523792, sep 2023. doi:10.1101/2023.01.12.523792. URL https://www.biorxiv.org/content/10.1101/2023.01.12.523792v3https://www.biorxiv.org/content/10.1101/2023.01.12.523792v3.abstract.

R. Wei, J. Wang, E. Jia, T. Chen, Y. Ni, and W. Jia. GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Computational Biology*, 14(1):e1005973, jan 2018. ISSN 15537358. doi:10.1371/journal.pcbi.1005973. URL https://doi.org/10.1371/journal.pcbi.1005973.g001.

S. Wieczorek, F. Combes, C. Lazar, Q. Giai Gianetto, L. Gatto, A. Dorffer, A.-M. Hesse, Y. Couté, M. Ferro, C. Bruley, and T. Burger. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, 33(1):135–136, jan 2017. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTW580. URL https://academic.oup.com/bioinformatics/article/33/1/135/2525689.

S. Wieczorek, F. Combes, H. Borges, and T. Burger. Protein-level statistical analysis of quantitative label-free proteomics data with ProStaR. *Methods in Molecular Biology*, 1959:225–246, 2019. ISSN 10643745. doi:10.1007/978-1-4939-9164-8_15. URL https://link.springer.com/protocol/10.1007/978-1-4939-9164-8_15.

Z. Zhou, C. Ye, J. Wang, and N. R. Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11(1):1–10, 2020. ISSN 20411723. doi:10.1038/s41467-020-14391-0. URL http://dx.doi.org/10.1038/s41467-020-14391-0.

R. A. Zubarev. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *PROTEOMICS*, 13(5):723–726, mar 2013. ISSN 1615-9861. doi:10.1002/PMIC.201200451. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pmic.201200451https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201200451https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.201200451.

A Supplementary Materials

A.1 PEMM limitations and comparison with Pirat

We present in this section some experiments illustrating PEMM's limitations. We rely on the original PEMM code, although the package is deprecated from the CRAN (as of 2022). First, we apply PEMM on all PGs of Ropers2021, with Γ parameters and K and λ hyperparameters obtained from the Pirat pipeline, as PEMM does not provide any way to tune them from the data. The other parameters are left as default. Among the 1830 PGs of the dataset, PEMM does not converge for 18% of them (a warning is returned in console), and the following error occurs in one PG: "Error in svd(X) : infinite or missing values in 'x'". In all non-converging cases, it happens because the likelihood becomes "NA" at some iteration. In addition, among these non-converging PGs, PEMM still returns 26% of the time a Σ matrix with a diagonal term exploding over 10^4 (up until 10^{50}). Moreover, a refined look at the behavior of the penalized likelihood computed by the PEMM function raises concern: On this dataset, it decreases in 16% of PGs. It does not necessarily happen when a PG does not converge; however, when it does, a Σ diagonal term explodes every time. We display on Figure 4 an example PG for which penalized likelihood decreases. Note that this behavior is in contradiction with the proof of convergence property of the penalized EM algorithm given in the original PEMM article. For all these reasons, PEMM is not included in Pirat's benchmark.

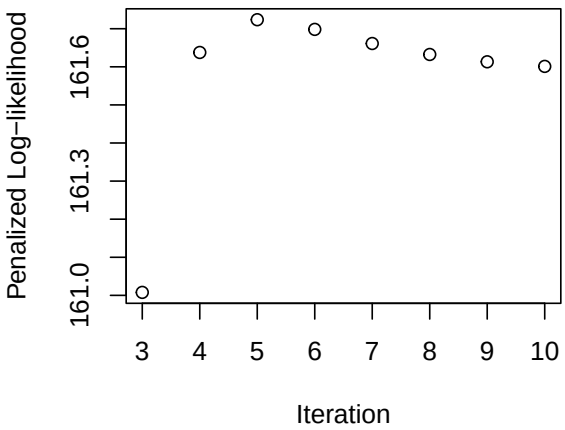


Figure 4: Penalized log-likelihood of PG 34 in Ropers2021 dataset, computed with `get.llk` function of PEMM package.

We claim that the convergence issue of PEMM relates directly to the analytical approximation used by the authors. In fact, a main issue regarding the estimation of μ and Σ by maximizing Equation 7, is that the expectation inside the log-likelihood is not analytically tractable for any missingness mechanism. It is why we choose in Pirat to derive its lower bound, which is provably tractable with PEMM's missingness mechanism (see Section 2.7), instead of maximizing it directly. This expectation also appears in PEMM's E-step. However, the author chose to compute it directly by approximating missingness mechanism Equation 5 as:

$$P(m_{i,j} = 1 | x_{i,j}, \Gamma) = \exp(-\gamma_0 - \gamma_1 x_{i,j}) \tag{18}$$

and by setting any observed value below $-\gamma_0/\gamma_1$ as missing. This approximation makes the log-likelihood unbounded and hinders convergence in some cases, according to the following argument. Using this approximation, we have (see Chen et al. [2014], Supp. Mat. Appendix D) $\forall i \in \{1, ..., n\}$:

$$\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}}, \Gamma)] = \exp(-\gamma_0|\{i, \text{mis}\}| - \gamma_1 \mathbf{1}^T \tilde{\mu}_i + 1/2 \cdot \gamma_1^2 \mathbf{1}^T \tilde{\Sigma}_i \mathbf{1}), \tag{19}$$

where $\tilde{\mu}_i, \tilde{\Sigma}_i$ are the parameters of the normal distribution $X_{i,\text{mis}}|X_{i,\text{obs}}$. The above quantity diverges towards $+\infty$ if any coefficient of $\tilde{\Sigma}_i$ tends towards $+\infty$, which is not consistent with the definition of a probability distribution. This analysis is consistent with the observed exploding variance terms during PEMM iterations in the previous experiment.

Beyond this essential improvement upon PEMM, Pirat differs on several other points that are listed in Table 2.

	PEMM	**Pirat**
Likelihood maximization method	Penalized EM algorithm	L-BFGS on the lower bound of log-likelihood
How to deal with untractable $\mathbb{E}_{X_{i,\text{mis}}\|.}[P(M_{i,\text{mis}}\|X_{i,\text{mis}},\Gamma)]$?	Approximate missingness mechanism by $\exp(-\gamma_0 - \gamma_1 x_{i,j})$	Find a tractable lower bound of it
(Hyper)parameters left to the user	λ, K, Γ	None

Table 2: Summary of main methodological differences between PEMM and Pirat.

Despite the practical limitations of PEMM, it is useful as a reference for Pirat evaluations, to check if its estimation is not hindered by relying on the lower bound of the penalized log-likelihood. To this end, we compare the mean parameter μ mean bias and the RMSE on synthetic data, where the missingness mechanism is controlled, while being identical to the one used in Pirat and PEMM algorithms. We use 1000 simulated datasets generated with 10 features and 10 samples, from a multivariate Gaussian distribution with a random mean distributing alike Ropers2021 and a Toeplitz covariance matrix with exponential decay. We apply a mask on each dataset using the model of Equation 5, with parameter Γ chosen such that about 15% of values are missing (alike Ropers2021), and with $\gamma_1 > 0$. We then apply PEMM and Pirat on each of them, providing both methods true Γ parameter. Finally, we compute the RMSE and mean bias between true and estimated feature mean for each dataset to subsequently display the corresponding box plots. As PEMM does not converge for 40% of the datasets generated (every time because of a "NA" likelihood), we only retain metrics of datasets for which PEMM converges (see 5). A Wilcoxon test between the distributions of the results of the two methods yields p-values of 0.03 (RMSE) and 0.12 (bias). To conclude, in this experimental setup where the missingness mechanism is known and provided and where the data leading to diverging estimates are ignored, Pirat estimate of μ is at least as good as that of PEMM.

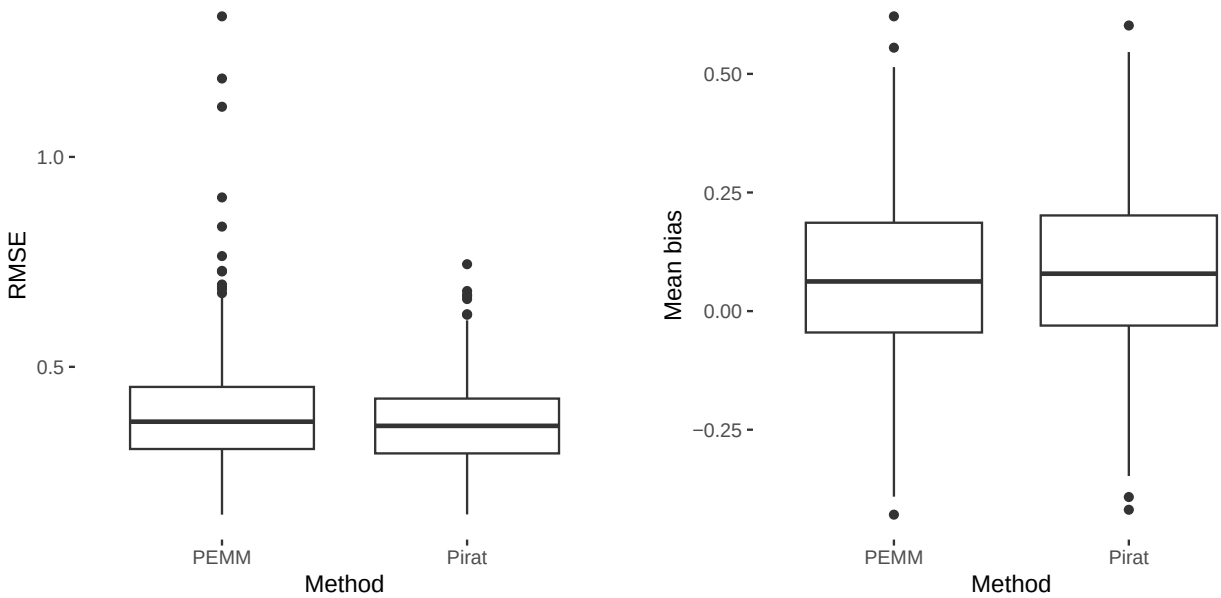


Figure 5: Boxplot of RMSEs between true and estimate μ among 1000 different synthetic datasets for PEMM and Pirat methods.

A.2 Dataset descriptions

We hereafter summarize the datasets used in our experiments, which are all publicly available, notably on ProteomeXchange. In parenthesis, we display for each of them the number of samples, peptides/precursors and PGs, respectively:

- **Cox2014** Cox et al. [2014] (8 samples, 22273 peptides, 2221 PGs): a mixture of 48 recombinant human proteins that are available as an equimolar mixture (UPS1) or mixed at defined ratios spanning 6 orders of magnitude (UPS2) with respect to UPS1 (10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}). UPS1 and UPS2 are separately

digested and resulting peptides are spiked into a *E. coli* lysate, resulting in two distinct experimental conditions. Each condition is analysed four times in data dependent acquisition (DDA, *i.e.*, a single precursor is iteratively selected for fragmentation and identification by the mass spectrometer according to its response level in the first MS acquisition), resulting in four technical replicates per conditions. We use the file `peptide.txt` available at repository PXD000279, containing peptide-level abundances.

- **Bouyssie2020** Bouyssié et al. [2020] (40 samples, 20975 peptides, 2219 PGs): a mixture of UPS1 proteins spiked at different concentrations in the same yeast lysate. The dataset contains 10 conditions with 4 technical replicates each (each condition corresponding to a different UPS1 quantity, namely, 0.01-0.05-0.1-0.250-0.5-1-5-10-25-50 fmol, for 1μg of yeast lysate). Hence, unlike Cox2014, within each condition, all the UPS proteins are spiked in with identical concentrations. Replicates are analysed in DDA mode. We use peptide intensities from `allsamples_sum.xlsx` file available at repository PXD009815.

- **Huang2020** Huang et al. [2020] (25 samples, 132140 precursors, 6454 PGs): the UPS2 mixture spiked in tissue lysates from 25 mouse cerebellum samples. Five conditions (with 5 technical replicates each) are generated from spiked UPS2 proteins in known and different concentrations (namely 0.75-0.83-1.07-2.04-7.54 amol/μl) into mouse cerebellum samples. The LC-MS/MS runs are acquired by the data independent method (DIA, *i.e*, a set of precursors in a given mass to charge range are iteratively selected for fragmentation and identification, in order to cover the whole MS acquisition, regardless of the measured intensities). We use the published Spectronaut results in `Spike-in-biol-var-OT-SN-Report.txt` available at repository PXD016647, and imputed MS1 quantification at precursor level.

- **Capizzi2022** Capizzi et al. [2022] (10 samples, 14256 peptides, 2616 PGs): a study on Huntington's disease effect on axonal growth in mouse. Two conditions with 5 biological replicates each are compared using LC-MS/MS in a data dependent acquisition mode (DDA), representing wild-type vs genetically modified model. Details of sample preparation can be found in the original article. The accession number for this dataset is PXD023885. Raw data is processed in the same manner as described in the original paper to obtain peptide-level quantification table.

- **Vilallongue2022** Vilallongue et al. [2022] (SCN:8 samples, 19050 precursors, 2741; SC:8 samples, 33364 precursors, 4238 PGs): a study on the influence of injury on visual targets in mouse. Five different tissue types are analysed by LC-MS/MS in DDA acquisition mode in independent quantification tables, with each time a control and an injured condition (4 biological replicates each). To test our method, we choose the two tissues that show the best and worse within-PG peptide correlations among the five (see 4.2): respectively, the suprachiasmatic nucleus tissue (SCN), and the superior colliculus (SC). The details of sample preparation can be found in the original article. The accession number for this dataset is PXD029325. Raw data is processed in the same manner as described in original paper to obtain precursor-level quantification table.

- **Habowski2020** Habowski et al. [2020] (18 samples, 26586 precursors, 3461 PGs): a study on the differentiation of mouse colon epithelial cells. Stem cells and 5 differentiated cells are analysed at proteomic and transcriptomic levels in an unpaired manner. Three biological replicates are used for each cell type in proteomic analyses. The transcriptome dataset is produced by Illumina paired-end stranded sequencing. The proteome dataset is produced by LC-MS/MS in DDA acquisition mode.

  - **Transcriptomic data processing**
    We download the raw sequencing data from GEO (GSE143915). Preprocessing and quality control are performed using the Trimmomatic Bolger et al. [2014] and FastQC tools respectively. Trimmed reads are aligned to the mouse GRCm39 genome assembly (filename given in Sup. Mat. Section A.3) by the STAR mapping software (version 2.7.8a) Dobin et al. [2013] provided with the Ensembl gene model given in Sup. Mat. Section A.3 for mapping splices. Read counts are generated using HTSeq (version 0.9.1; option -s no) Anders et al. [2015]. DESeq2 version 1.22.1 Love et al. [2014] is used to generate normalized counts.

  - **Proteomic data processing**
    We download the raw spectra from ProteomeXchange (accession ID: PXD019351). Peptides and proteins are identified using Mascot (version 2.7.0.1, Matrix Science) searching the Ensembl protein database for *Mus musculus* GRCm39 (filename given in Sup. Mat. Section A.3) appended with an in-house classical contaminant database. Mascot search is performed with the following parameters: trypsin/P as enzyme and two missed cleavages allowed; precursor and fragment mass error tolerances at 10 ppm and 20 ppm, respectively. The following peptide modifications are allowed during the search: Acetyl (Protein N-term, variable) and Oxidation (M, variable). The Proline software Bouyssié et al. [2020] is used to validate identifications: conservation of rank 1 peptides, peptide length ≥ 7

amino acids, false discovery rate (FDR) of peptide-spectrum-match identifications $< 1\%$ as calculated by Benjamini-Hochberg procedure and minimum of 1 specific peptide per identified protein group. Peptide ion intensities are calculated from extracted mass spectrum intensities of all peptides and normalized using variance stabilizing transformation with Prostar Wieczorek et al. [2017]. A more detailed description of the parameters used in the quantification step are provided in Supplementary Table 1.

- **Ropers2021** Ropers et al. [2021] (18 samples, 17472 precursors, 1830 PGs): a study on growth-arrested and wild type *E. coli* cells carrying a plasmid for glycerol production at various time points. A wild type and a modified strain are analysed in a paired manner at proteomic and transcriptomic levels. The modified strain is analysed at four different time points, resulting in four different conditions, and two others from the wild-type on the two first time points. Three biological replicates are made for each condition. The transcriptome dataset is produced by stranded sequencing on the Ion S5 using the Ion 540 chip. The proteome dataset is produced using LS-MS/MS in DDA acquisition model.

  - **Transcriptomic data processing**: Same procedure than for Habowski2020 is applied. GEO accession number is GSE168336. Trimmed reads are aligned to the reference *E. coli* K12 substrain MG1655 genome (Genbank assembly accession: GCA_000005845) provided with the Ensembl gene model given in Sup. Mat. Section A.3 for mapping splices.
  - **Proteomic data processing**:

    The same procedure than for Habowski2020 is applied, except for the following: Raw spectra are downloaded from ProteomeXchange (accession ID: PXD024231). Peptides and proteins are identified by searching the same Ensembl protein database used in transcriptomic analysis (for this Ensembl genome assembly of *E. coli* the protein identifier is identical to the transcript identifier), appended with an in-house classical contaminant database, the plasmid pCL1920 protein sequences and the protein sequences of the GPD1 and GPP2 yeast genes which were cloned into it (Uniprot accessions: Q00055 and P40106). Carbamidomethyl (C, fixed) peptide modification is additionaly allowed during search. Peptides with length ≥ 6 amino acids are conserved at validation of identifications. A more detailed description of the parameters used in the quantification step are provided in Supplementary Table 2.

A.3 Ensembl Gene models, genome assembly, protein databases

This section refers to the filenames of Ensembl gene model, genome assembly and protein databases used to build proteomic and transcriptomic datasets of Ropers2021 and Habowski2020 (see Section 3.1).

A.3.1 Ropers2021

The Ensembl genome assembly used for read alignment in transcriptomic analysis is the following:
`GCA_000005845.2.fasta`

The Ensembl gene model used for mapping in transcriptomic analysis and for peptide identification in proteomic alignment is the following:
`Escherichia_coli_str_k_12_substr_mg1655_gca_000005845.ASM584v2.51.gtf`

A.3.2 Habowski2020

The Ensembl genome assembly used for read alignment in transcriptomic analysis is the following:
`Mus_musculus.GRCm39.dna.primary_assembly.fa`

The Ensembl gene model used for mapping slices in transcriptomic analysis is the following:
`Mus_musculus.GRCm39.104.gtf`

The Ensembl protein database used for peptide identification in proteomic analysis is the following:
`Mus_musculus.GRCm39.pep.all.fa`

A.4 ROC curves

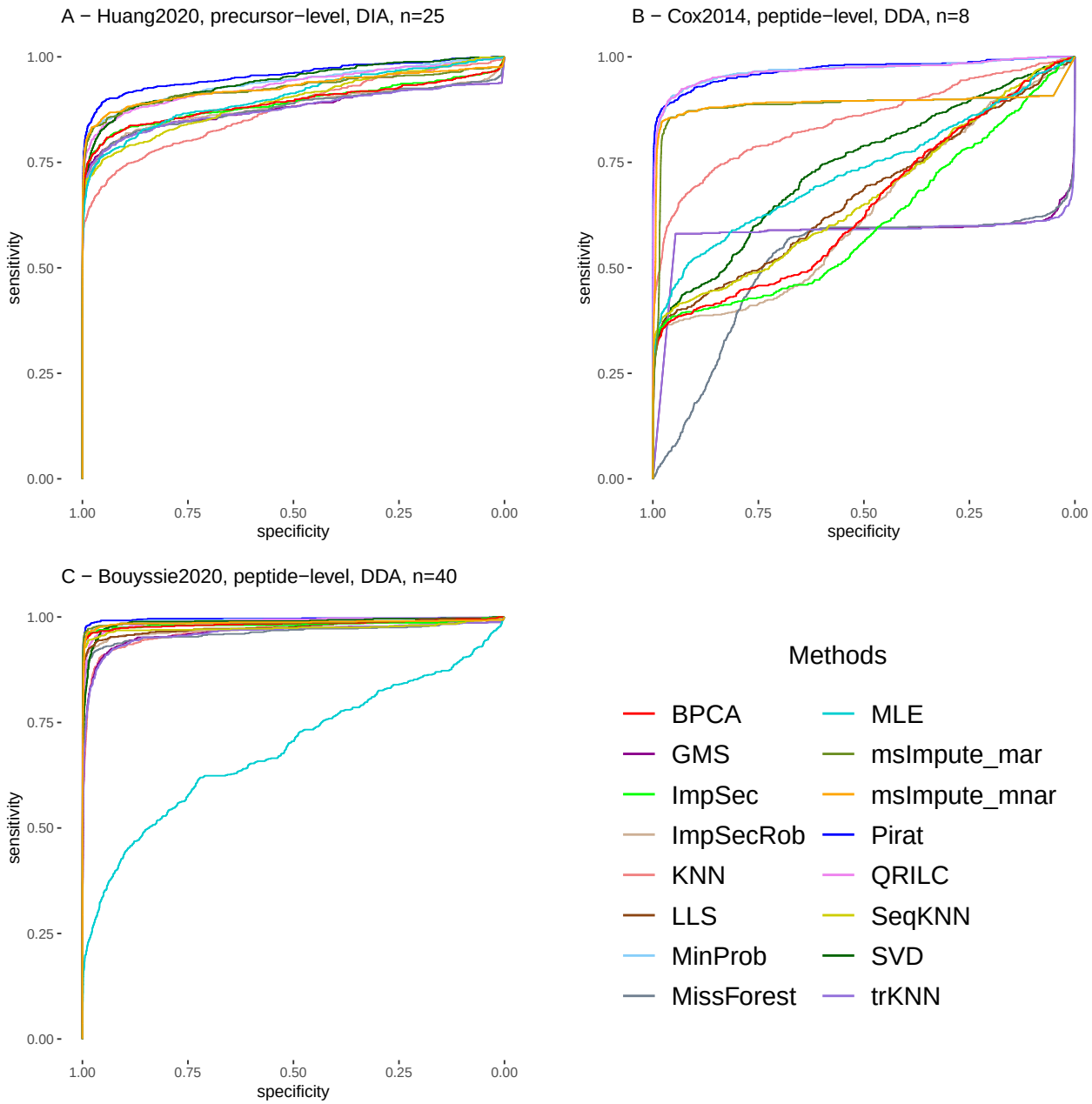


Figure 6: Complete ROC curves from p-values associated to the differential analysis validation on the three datasets Huang2020, Cox2014 and Bouyssié2020.

A.5 Mean MAE for Capizzi2022 and Vilallongue2022 with best methods

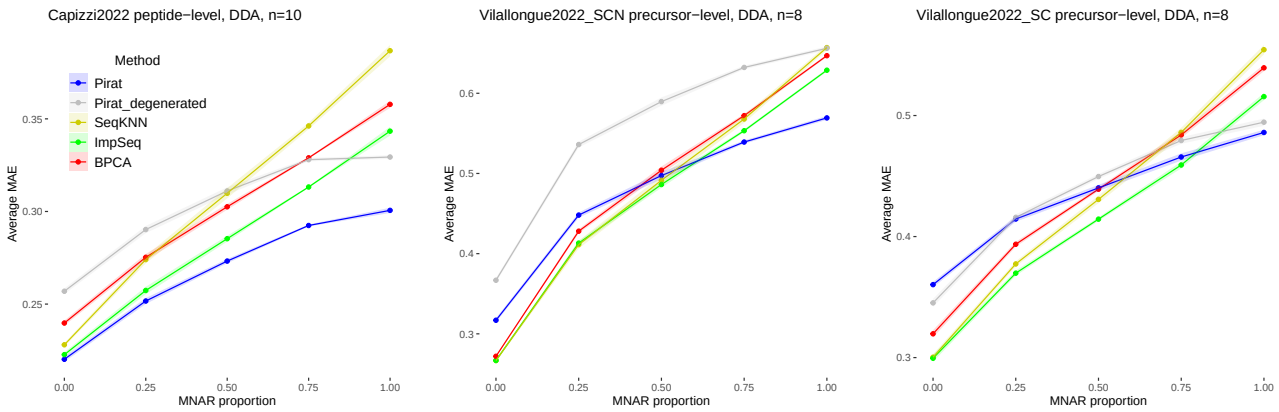


Figure 7: Average MAE (bottom) of BPCA, ImpSeq, Pirat, Pirat-Degenerated and SeqKNN in function of the proportion of MNAR values on Capizzi2022 (left), Villalongue2022 SCN (center) and Villalongue2022 SC. The imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are also indicated. The average MAE was computed over 5 different seeds, and margins correspond to standard deviation.

A.6 Mean MAE and RMSE with QRILC and MinProb

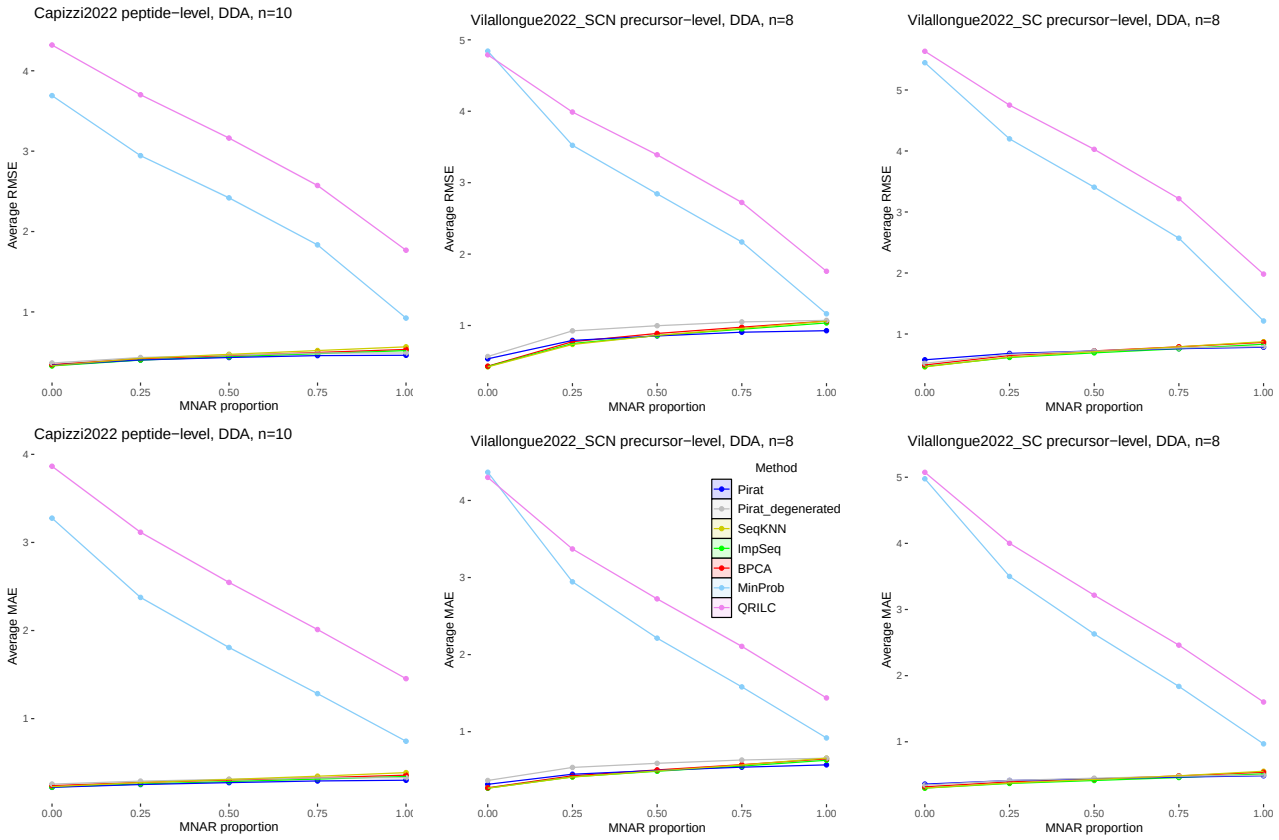


Figure 8: Average RMSE (top) and MAE (bottom) of MinProb, QRILC, BPCA, ImpSeq, Pirat, Pirat-Degenerated and SeqKNN in function of the proportion of MNAR values on Capizzi2022 (left), Villalongue2022 SCN (center) and Villalongue2022 SC. The imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are also indicated. The average RMSE and MAE was computed over 5 different seeds, and margins correspond to standard deviation.

A.7 Quantitative two-peptide rule (Pirat-2)

Even though it is technically possible, one should wonder whether it is reasonable to impute missing values for singleton PGs. First, protein abundances for those PGs will directly result from imputation. Consequently, the quality of the imputation will weight a lot in downstream analysis (*e.g.*, differential analysis for biomarker discovery). Second, we cannot exploit the peptide-wise covariances for imputation, and thus suspect deterioration of the imputation for singleton PGs. To confirm this, we compare the median imputation error on singleton PGs with the error of PGs of size greater than two in the MNAR and MCAR settings. Results from this experiment (Figure 9) show that in most cases MVs from singleton PGs are indeed harder to impute than for all other PGs. Akin to the two-peptide rule in protein identification, a simple but effective strategy is to filter out singleton PGs that require missing value imputation to avoid errors in the imputation step to excessively affect downstream analysis. We call this strategy the *quantitative two-peptide rule* or simply the *two-peptide rule*.

Looking more closely at Figure 9, the results of Habowski2020 in the MNAR setting are intriguing: we do not observe that the pseudo-MNAR values in this dataset are harder to impute on singleton PGs than on other PGs. A plausible explanation is that the pseudo-MVs of non-singleton peptides (which we compute the error on) in this setting are located on peptides containing mostly MVs. To confirm this, we show in Sup. Mat. Figure 10 (A, B, C, D) the histograms of the number of MVs contained in non-singleton-peptides carrying pseudo-MVs. We see that histogram A, representing Habowski2020 MNARs, has a clearly different trend from other scenarios: It is the only situation where pseudo-MVs are, for the vast majority, carried by very scarce peptides containing mostly MECs. Hence, correlations estimated by Pirat are uninformative as they are based on very few values coming from other conditions, which concurs with the drop of performances over non-singleton PGs with respect to singleton PGs in the MNAR setting.

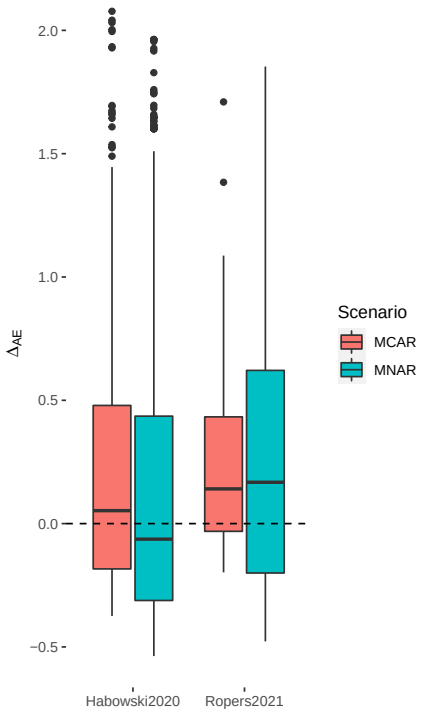


Figure 9: Boxplots of the distributions of the differences (denoted as Δ_{AE}) between (i) the absolute errors in singleton PGs and (ii) the median of the absolute errors in all other PGs, after Pirat imputation, for a given dataset (Habowski202 or Ropers2021) and MNAR / MCAR scenario. Hence, for a given dataset and MNAR setting, the part of the boxplot that is above zero corresponds to absolute errors in singleton PGs that are greater than the median of absolute errors for non-singleton PGs.

A.8 MV distribution in Habowski2020 and Ropers2021 experiments

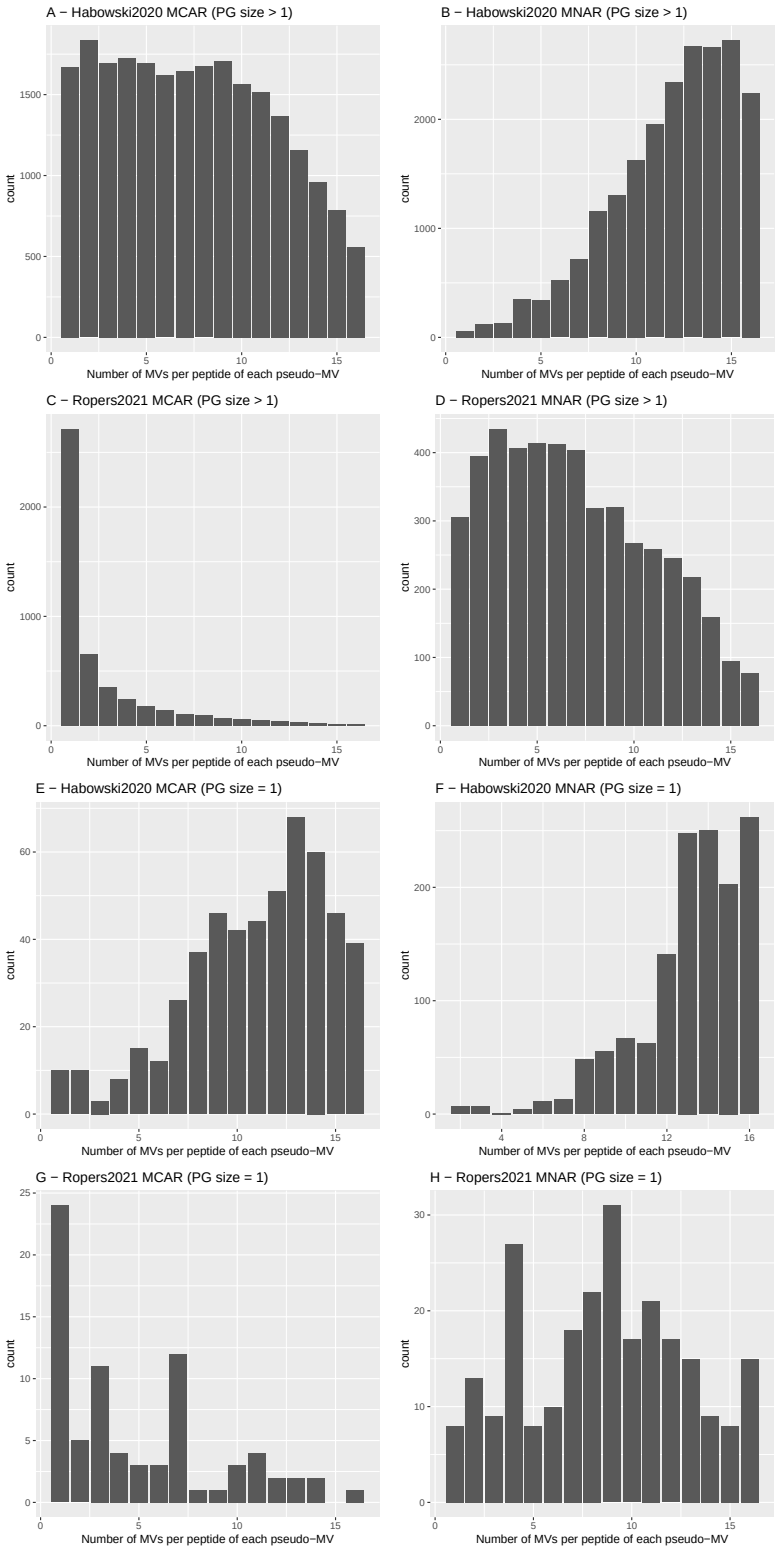


Figure 10: Histograms counting, for each pseudo-MV, the number of MVs (real or pseudo) in the same peptide, for Habowski2020 (A, B, E, F) and Ropers2021 (C, D, G, H), in MCAR (A, C, E, G) and MNAR (B, D, F, H) setting, and over 10 different seeds. We separate histograms for peptides contained in singleton PG (A, B, C, D) and non-singleton PGs (E, F, G, H). Note that the number of samples equals 18 in both datasets.

A.9 Fitting of missingness mechanism

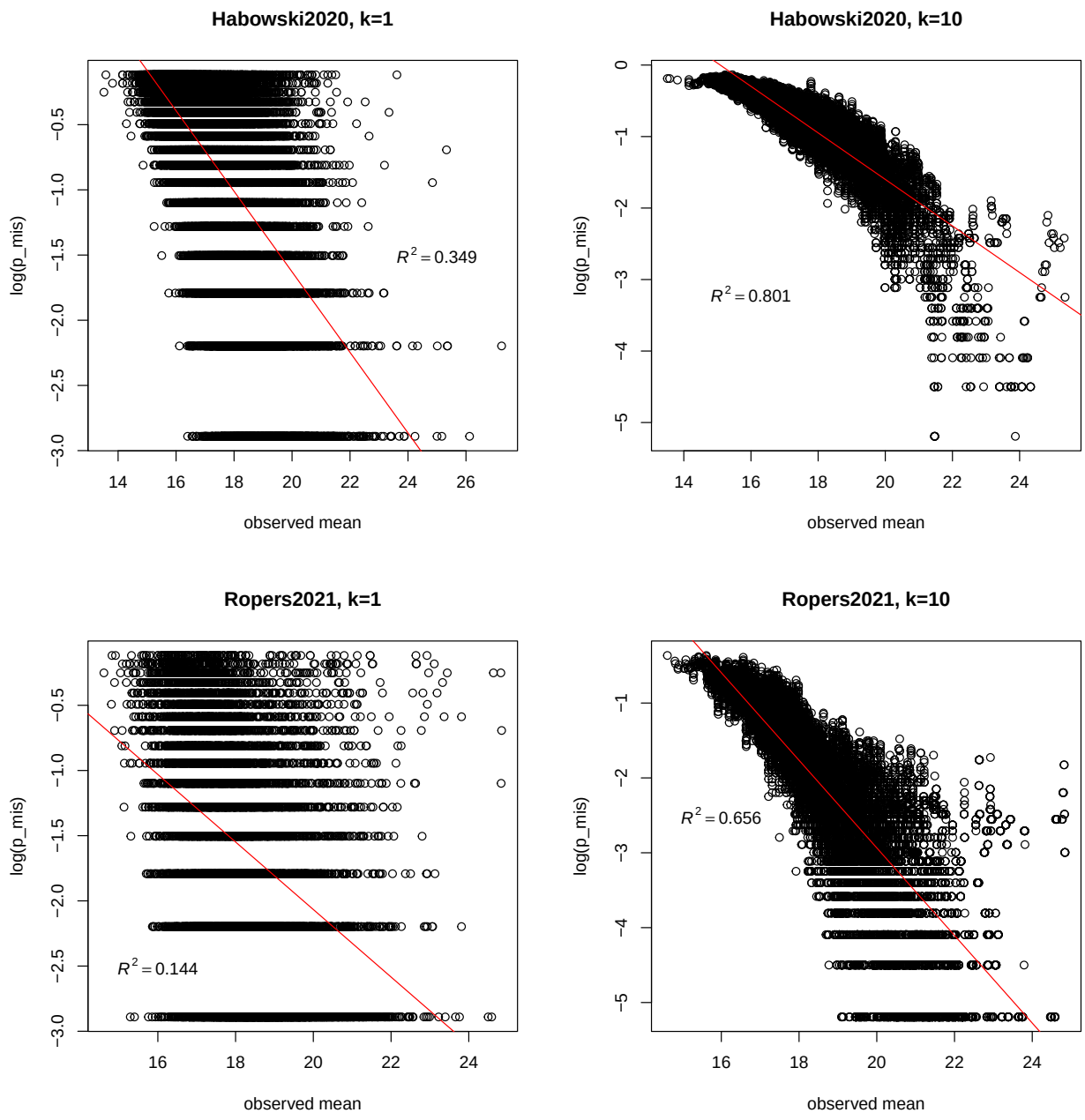
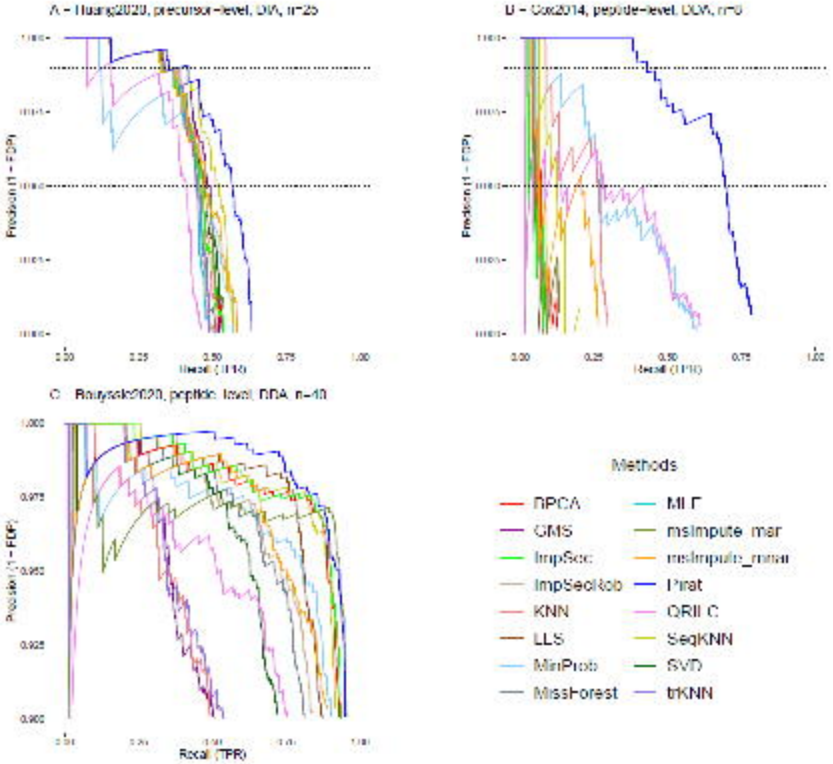
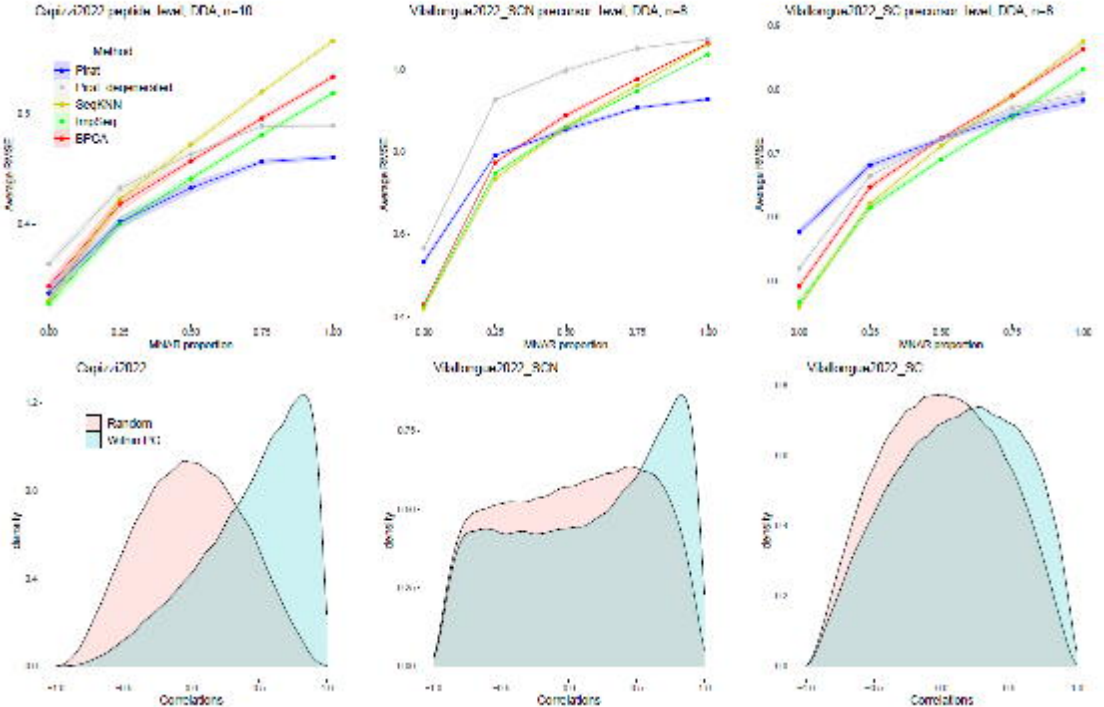
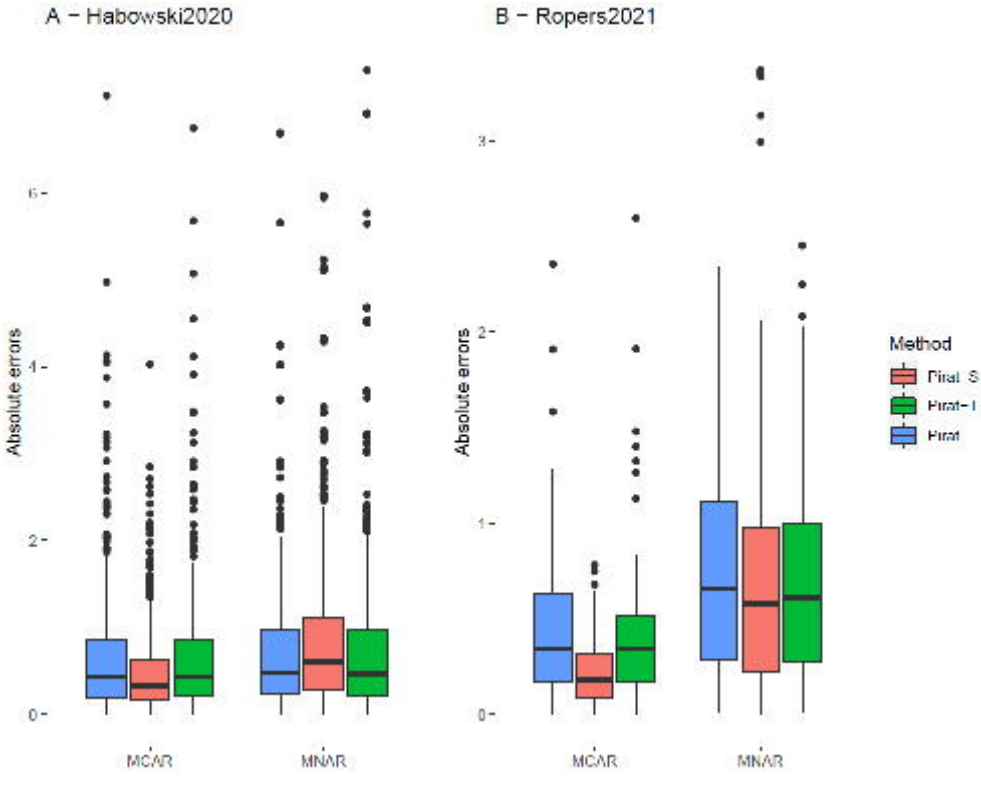


Figure 11: Regression of the log-probability of missing onto mean observed abundance following method described in 2.4, for Habowski2020 and Ropers2021, for $k = 1$ and $k = 10$. Residuals sum of squares (R^2) of the linear regression are also displayed.


A – Huang2020, precursor-level, DIA, n=25
B – Cox2014, peptide-level, DDA, n=6
C – Bouyssié2020, peptide-level, DDA, n=40
Precision (1 – FDP)
Recall (TPR)
Methods
DPCA
GMS
ImpSeq
ImpSeqRob
KNN
LLS
MinProb
MissForest
MLE
msImpute_mar
msImpute_mnar
Pirat
QRILC
SeqKNN
SVD
trKNN


Capizzi2022 peptide level, DDA, n=10
Method
Pirat
Pirat-degenerated
SeqKNN
ImpSeq
EPCA
Average RMSE
MNAR proportion
Vitalongue2022_SCN precursor level, DDA, n=8
Vitalongue2022_SC precursor level, DDA, n=8
Capizzi2022
Random
Vitalongue2022_SCN
Vitalongue2022_SC
density
Correlations


A – Habowski2020
B – Ropers2021
Absolute errors
0
2
4
6
1
3
MCAR
MNAR
Method
Final S
Final-I
Final