Journal of Extracellular Biology

RESEARCH ARTICLE OPEN ACCESS

BioProEV: A Bioinformatics Pipeline for Biologically-Relevant Handling of Missing Values in the Analysis of Extracellular Vesicles by Mass Spectrometry

Pâmella Miranda[1,2] | Jose G. Marchan-Alvarez[1,2] | Annemarijn Offens[3,4] | Ruihan Zhou[1,2] | Mathieu Y. Brunet[1,2] | Loes Teeuwen[3,4] | Maria Eldh[3,4] | Susanne Gabrielsson[3,4] | Phillip T. Newton[1,2]

[1]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden | [2]Astrid Lindgren Children's Hospital, Stockholm, Sweden | [3]Division of Immunology and Allergy, Department of Medicine (Solna), Karolinska Institutet, Stockholm, Sweden | [4]Clinical Immunology and Transfusion Medicine, Karolinska University Hospital, Stockholm, Sweden

Correspondence: Phillip T. Newton (phillip.newton@ki.se)

Received: 19 January 2026 | **Revised:** 31 March 2026 | **Accepted:** 29 April 2026

Keywords: Bovine milk | extracellular vesicles | fetal bovine serum (FBS) | missing values | proteomics

ABSTRACT

While proteomics is being increasingly applied to investigate extracellular vesicles (EVs), there remains no consensus on how to address the inevitable missing values within proteomics datasets. Here, we devised a three-step approach that prioritized retaining biologically-relevant information in EV samples using a dataset containing two populations of EVs analysed by liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS). Firstly, to avoid overfitting, we excluded proteins in which more than half of the values were missing. Next, we statistically tested for low protein abundance in a single population: when the number of potential "missing not at random" (MNAR) values was significantly enriched, these values were replaced with the lowest possible value of "1". Finally, all remaining missing values were then imputed using the Random Forest machine learning algorithm. Our final dataset included 49.9 % of proteins that originally contained at least one missing value, from which 84.7 % were listed in the ExoCarta database, significantly more than would be expected by chance, strongly indicating biologically relevant imputation. To enable other EV researchers to analyse proteomics data in a robust, easy-to-use and peer-reviewed manner, we provide BioProEV, a bioinformatics pipeline to impute missing values with biological relevance in EV datasets.

1 | Introduction

Extracellular vesicles (EVs) are small membrane-bound particles, released by virtually all cells into the extracellular space, that are increasingly recognized as important mediators of inter-cellular communication since they contain biologically active molecules including nucleic acids, proteins and lipids. Among these, proteins, both internal and surface-related, are integral components of EVs that can determine their properties and functions, and therefore, important information can be obtained by analysing the protein compositions of EV preparations (Iraci et al. 2016; Yáñez-Mó et al., 2015). For example, some proteins are typically enriched in EVs independently of their cellular origin; these proteins, such as the tetraspanins (e.g. CD63, CD81 and CD9), and cytosolic proteins (e.g. Annexins and PDCD6IP [also known as ALIX]), are often used to ascertain whether or not isolated nanoparticles can be identified as EVs (Welsh et al. 2024). Additionally, certain proteins present in EVs reflecting their cell of origin, such that EVs obtained from biological fluids like blood and urine, can be used diagnostically or prognostically

to identify pathological changes (Kumar et al. 2024). For example, circulating EVs enriched in cysteine-rich protein 61 (Cyr61), human ether-a-go-go-related gene 1 (hERG1) and heat shock protein 47 (Hsp47) have predictive potential for cardiovascular diseases (W. Li et al. 2021; Osorio et al. 2024), while other proteins such as alpha 1-antitrypsin, Histone H2B type 1-K (H2B1K) and Tumour-associated calcium signal transducer 2 (TACSTD2) in urinary EVs have been identified as diagnostic and prognostic biomarkers for bladder cancers (Chen et al. 2012; Lin et al. 2016). Furthermore, other proteins can confer functional properties to EVs under both physiological and pathological conditions, such as the modulation of immune function (Kalluri 2024) or the regulation of mineralization of extracellular matrix during skeletal growth (Cui et al. 2016), as well as stimulating liver metastases (H. Zhang et al. 2017) or promoting atherosclerotic plaque reactivity (Oggero et al. 2021). Hence, proteomics is becoming an increasingly important tool for EV researchers to explore the roles of EVs in both physiological and pathological settings, and robust approaches to analyse proteomics data are needed to allow meaningful interpretation and reproducibility.

Missing values are absences of data within a dataset in which individual samples lack a numerical score for a certain protein, and are inevitable in proteomics datasets (Karpievitch et al. 2012; Karpievitch et al. 2009; Webb-Robertson et al. 2015). While the number of missing values in proteomics data varies, it is usually greater than 50 % of the total data, with at least 50 % of the given proteins (variables) likely to have at least one missing value (Albrecht et al. 2010; Karpievitch et al. 2009; Lazar et al. 2016; Liu and Dongre 2021). This absence in the data can greatly affect the downstream analyses, reducing the power of statistical tests and leading to bias in the interpretation of results (Albrecht et al. 2010; Jin et al., 2021; Zhou et al. 2018). Detection and quantification of EV proteins is more challenging than cell-derived proteins, as no available method can purify EVs completely and different purification technologies could result in different proteomic analysis results. Preparations from biological fluids with lower EV concentrations are likely to contain relatively more contaminating non-EV-related proteins than those with higher EV concentrations, leading to an increased chance of missing values across samples. Furthermore, normalization of input material based on original bodily fluid volume, protein concentration or particle numbers all have their challenges and limitations. Despite this, missing value handling is frequently not addressed in methodological descriptions within the EV literature, compromising the reliability and reproducibility of published studies (Karimi et al. 2018; Kowal et al. 2016; Kreimer et al. 2015; Singh et al. 2025).

Missing values are caused by biological and analytical factors, including the absence of a protein from a sample, protein abundance below the instrument's detection limit, sample loss in preparation, miscleavage of peptides during the digestion process, and instrument calibration (Albrecht et al. 2010; Jin et al., 2021; Karpievitch et al. 2012; Zhou et al. 2018). The missing values can be classified by three missingness mechanisms: whereas missing at random (MAR, which depends only on the observed data and is related to sample preparation or instrument settings) and missing completely at random (MCAR, which does not depend on either observed or missing data) both appear randomly in the dataset, missing not at random (MNAR, which depends only on the missing data) is caused by relatively low protein abundance for biological reasons (Little and Rubin, 2019; RUBIN 1976). In proteomics data, missing values usually represent a mixture of MNAR (protein abundance-dependence) and MCAR/MAR (protein abundance-independence) (Albrecht et al. 2010; Jin et al., 2021; Karpievitch et al. 2012; Webb-Robertson et al. 2015).

Since there is no rule of thumb for handling missing data (Jin et al., 2021; Webb-Robertson et al. 2015), we developed a novel approach to investigate proteomics data obtained from two populations of EVs. Thus, we aimed to devise a method to handle missing values in order to maintain meaningful biological differences in the EV data while minimizing the influence of imputation methods to introduce spurious results. We combined these steps into a bioinformatics pipeline, named BioProEV, which we applied to two EV populations, fetal bovine serum (FBS) and bovine milk EVs (each analysed in triplicate), obtained from liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS).

2 | Methods

2.1 | Standardization of Experimental Parameters of EV Studies

To contribute to transparency and reproducibility in the EV field, all experimental details have been submitted to the EV-TRACK knowledgebase (Van Deun et al. 2017) under the following accession numbers: EV250092 for FBS-derived EVs and EV250104 for milk-derived EVs.

2.2 | Isolation of FBS-Derived EVs (FBS-EVs)

Heat-inactivated fetal bovine serum (FBS; Gibco, Cat. No. 10082147) was diluted to 30% (v/v) in DMEM/F-12 medium supplemented with GlutaMAX (Gibco, Cat. No. 31331093), followed by sequential centrifugation steps to remove potential protein aggregates and larger vesicles (Figure 1A). The diluted FBS was first centrifuged at 300×g for 5 min at 4°C, then at 3,000×g for 30 min at 4°C, and subsequently at 10,000×g for 30 min at 4°C. The resulting supernatant was passed through 0.22 μm bottle-top vacuum filters (Nalgene Rapid-Flow, Cat. No. 568-0020). Next, the filtered supernatant was aliquoted in ultra-clear tubes (Beckman Coulter, catalogue number: 344059) and subjected to ultracentrifugation at 100,000×g for 1 h at 4°C using the Optima L-90K ultracentrifuge (SW 40 Ti rotor). The EV-containing pellet was gently resuspended in 0.22 μm-filtered phosphate-buffered saline (PBS, 1X), and a second ultracentrifugation step was performed under the same conditions for 2 h. The final pellet was resuspended in 500 μL of 0.22 μm-filtered PBS (1X) and purified using size-exclusion chromatography (SEC) as previously described (Marchan-Alvarez et al. 2024). Briefly, the resuspended EVs were loaded onto Gen 2 SEC columns (70 nm pore size; Izon Science), and fractions 7 to 10, corresponding to the EV-enriched eluate, were collected and pooled (~2 mL) according to the manufacturer's protocol. The pooled fractions were concentrated to ~100 μL using Vivaspin 2 centrifugal filters

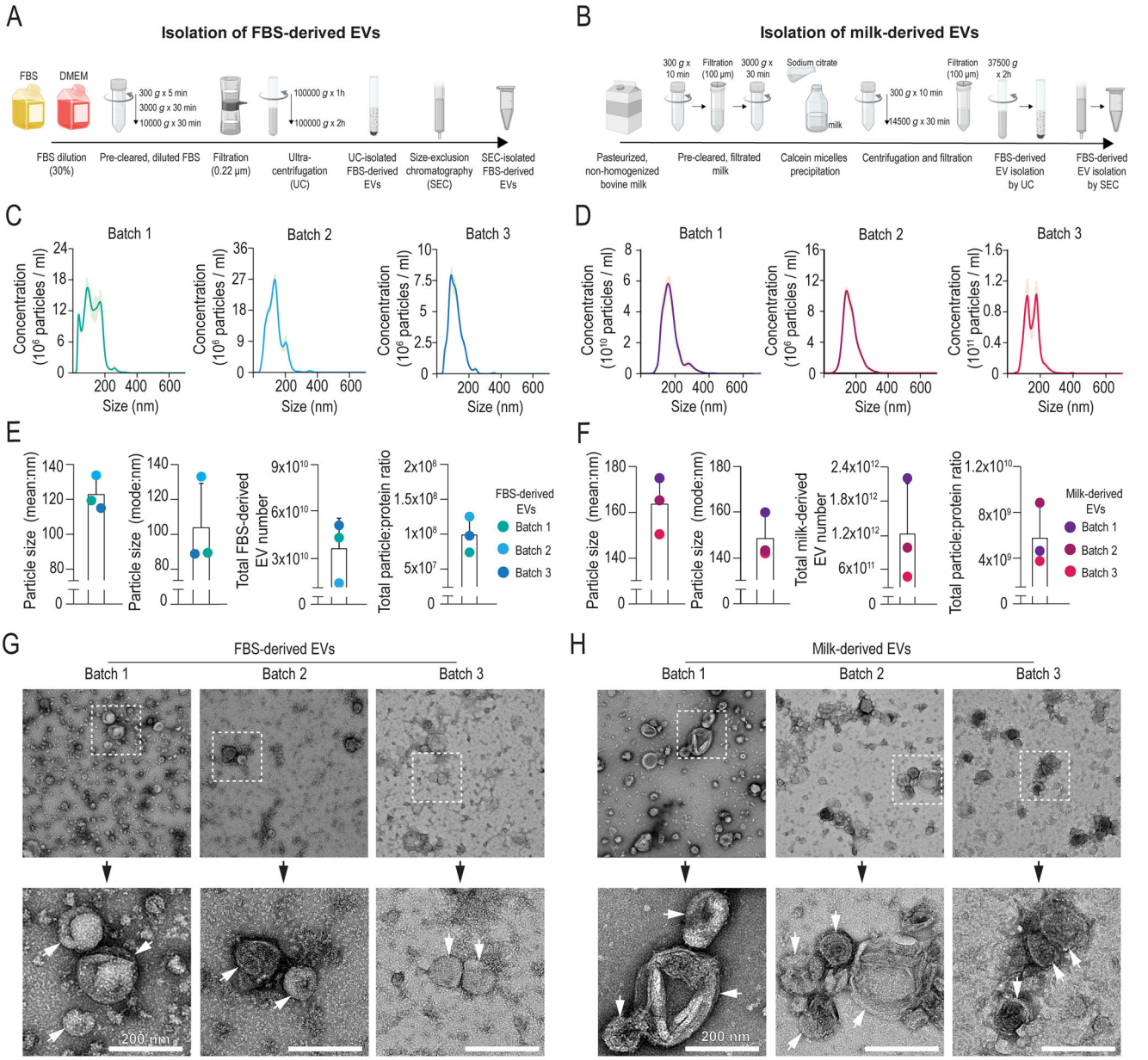


FIGURE 1 | Isolation and characterisation of FBS- and milk-derived EVs.

A. Schematic overview of the FBS-derived EV isolation workflow. Fetal bovine serum (FBS) was diluted with DMEM/F12 medium supplemented with GlutaMAX, followed by sequential centrifugation to remove debris. The supernatant was then filtered and subjected to ultracentrifugation. The resulting EV pellet was further purified by size-exclusion chromatography (SEC). **B**. Schematic overview of the milk-derived EV isolation workflow. Commercial pasteurised milk was processed through sequential centrifugation and filtration, followed by calcein micellae precipitation using sodium citrate (2%). The resulting suspension was centrifuged and filtered again, then ultracentrifuged. EVs were obtained from the final pellet by SEC. C–D. Nanoparticle tracking analysis (NTA) of FBS-derived EVs (left) and milk-derived EVs (right), showing particle concentration and size distribution. **E**–**F**. Quantitative analysis of NTA data for FBS-derived EVs (left) and milk-derived EVs (right), including mean and mode particle size, particle concentration, and particle-to-protein ratio. **G**–**H**. Representative transmission electron microscopy (TEM) images of FBS- and milk-derived EVs.

with a 10 kDa molecular weight cut-off (Merck, Cat. No. Z614238). Final EV preparations were stored in 0.22 µm-filtered PBS (1X) using low-retention polypropylene microcentrifuge tubes (Axygen Maxymum Recovery, Corning, Cat. No. MCT-150-L-C). Three individual preparations of FBS-EVs, referred to as "batches", were prepared for analysis.

2.3 | Isolation of Milk-Derived EVs (milk-EVs)

EVs were isolated from undiluted, pasteurized (72°C for 15s), non-homogenized bovine milk (Arla lantmjölk, 500 mL per isolation) purchased on the day of processing (Figure 1B). After gentle homogenization by inversion, milk was centrifuged at 300×g

for 10 min at 4°C to remove cellular debris, and the resulting cream layer was carefully displaced to allow collection of the underlying milk. The supernatant was filtered through a 100 μm cell strainer (Falcon. Cat. No. 352360) and subjected to a second centrifugation at 3,000×g for 30 min at 4°C, followed by cream layer removal as described above. The clarified milk was then mixed 1:1 (v/v) with a freshly prepared 4 % sodium citrate solution (Merck, Cat. No. 1.06448) to reach a final concentration of 2 %, and incubated on a gentle rocker at 4°C for 15 min. After repeating the 300×g and 3,000×g centrifugation steps, the supernatant was transferred into ultracentrifuge tubes (Beckman Coulter. Cat. No. 355622) and ultracentrifuged at 16,500×g for 30 min at 4°C using a Beckman Coulter ultracentrifuge (Type 45 Ti rotor). The supernatant was further ultracentrifuged at 100,000×g for 2 h at 4°C to pellet EVs. The resulting pellets were resuspended in PBS (1X) using gentle pipetting and strained through a 100 μm cell strainer. Two additional washing steps were performed in PBS with intermittent straining. Final EV suspensions were prepared in 15–30 mL of PBS, filtered again using a 100 μm cell strainer, and rotated overnight at 4°C to ensure complete dispersion. EV preparations were stored at −80°C until further analysis.

Following ultracentrifugation-based isolation, at least 3 mL of milk-derived EVs or a complete EV batch deriving from 80–160 mL of milk was subjected to further processing. To minimize filter clogging by large aggregates, samples were first centrifuged at 300×g for 2 min at 4°C. The milk-derived EV preparations were sequentially filtered through 0.8 μm and 0.22 μm sterile filters, pre-wetted with PBS. For size-based purification, samples were applied to qEVoriginal size-exclusion chromatography columns (70 nm; Izon Science). EV-enriched fractions (fractions 7–10) were collected in 2 mL microcentrifuge tubes. To concentrate the EV-containing fractions, 30 kDa molecular weight cutoff centrifugal filters (Amicon. Cat. No. UFC9030) were preconditioned by spinning PBS (1X) at 3,000×g until the membrane was equilibrated. SEC fractions were then loaded and centrifuged at 3,000×g for 8 min at 4°C. Multiple spins were typically required to reduce the volume to 700–900 μL for mEVs, while flow-through fractions concentrated more rapidly. The EVs were recovered by pipetting 100 μL of PBS along the filter membrane to elute residual vesicles adhered to the surface. The final concentrated EVs were transferred to cryovials (Sarstedt. Cat. No. 72.377 or 72.379) and stored at -80°C for downstream applications. Three individual preparations of milk-EVs, referred to as "batches", were prepared for analysis.

2.4 | Nanoparticle Tracking Analysis (NTA)

Particle concentration and size distribution of each EV preparation were assessed using NTA as described previously (Marchan-Alvarez et al., 2024). Samples were diluted in 0.22 μm-filtered PBS (1X) and analysed using a NanoSight LM10-HS instrument (Malvern Panalytical Ltd) equipped with a 488 nm laser and NTA 3.0 software. For each sample, five 30-second videos were captured in light scatter mode under consistent camera settings (level 13), using a syringe pump at a flow rate of 50, all at room temperature. Analysis parameters were kept constant across all measurements, including a screen gain of 10 and a detection threshold of 3.

2.5 | Protein Quantification in FBS- and Milk-Derived EV Preparations

Protein concentration for the FBS-derived EVs was quantified using the Micro BCA Protein Assay Kit (Thermo Scientific. Cat. No. 23235) as previously reported (Marchan-Alvarez et al., 2024). In brief, 5 μL of EV sample was mixed with 145 μL of the Micro BCA working reagent. Next, a standard curve was generated using serial dilutions of the albumin standard supplied by the manufacturer to determine protein concentrations. Absorbance was measured at 562 nm using a spectrophotometer (BMG Labtech, FLUOstar omega). The particle-to-protein ratio was then calculated by dividing the EV particle count obtained via NTA by the corresponding protein concentration measured from freshly isolated EVs. The protein concentration of the milk-derived EV preparations was assessed using DC protein assay (BioRad. Cat. No. 5000112) following the manufacturer's recommendations. Briefly, 20 μL of (diluted) EV sample or protein standard was mixed with 10 μL of reagent A, after which 80 μL of reagent B was added. The plate was read at 650 nm.

2.6 | Transmission Electron Microscopy (TEM) of FBS- and Milk-Derived EVs

EVs were prepared for TEM using negative staining as we previously reported (Marchan-Alvarez et al., 2024). Briefly, 5 μL of each EV sample was applied to a glow-discharged, carbon-coated grid and incubated for 1 min. Excess liquid was gently blotted with filter paper, followed by a 7-second incubation with 5 μL of 1 % uranyl acetate in water for contrast enhancement. After removing the excess stain with filter paper, the grids were allowed to air dry. Imaging was performed using a Talos 120C G2 transmission electron microscope (Thermo Scientific) operating at 120 kV and equipped with a CETA-D camera.

2.7 | Liquid Chromatography Electrospray Ionization Tandem Mass Spectrometry (LC-ESI-MS/MS)

Online LC-MS analysis was conducted using a Dionex UltiMate 3000 RSLCnano System coupled to an Orbitrap Exploris 480 mass spectrometer (Thermo Scientific). An aliquot of approximately 2 μg of each sample was suspended in LC mobile phase A. Equal peptide amounts derived from each EV sample (FBS- and milk-derived) were injected into the LC–MS/MS system to minimize technical variability; this approach ensures that differences observed across samples reflect biological variation rather than disparities in sample concentration. Peptides were first trapped on a C18 guard desalting column (Acclaim PepMap 100, 75 μm × 2 cm, nanoViper, C18, 5 μm, 100 Å) and subsequently separated on a 25 cm C18 column (Aurora Ultimate, C18, 1.7 μm, 25 cm × 75 μm). The mobile phase (nano capillary) consisted of solvent A (0.1 % formic acid in water) and solvent B (0.1 % formic acid in acetonitrile). Using a constant flow rate of 0.25 μL min^{-1}, a gradient was applied, starting at 6 % B and increasing to 43 % B over 180 min, followed by a sharp rise to 100 % B over 5 min.

Data acquisition was performed with FTMS master scans at a resolution of 120,000 (400–1200 m/z). Data-dependent MS/MS

(60,000 resolutions) was performed on the top five most abundant ions using higher energy collision dissociation (HCD) with a normalized collision energy of 30 %. Precursor ions were isolated with a 2 m/z window. Automatic gain control (AGC) targets were set to 1×10^6 for MS1 and 2×10^5 for MS2, with maximum injection times of 100 ms for both MS1 and MS2. The total duty cycle duration was approximately 2.5 s. Dynamic exclusion was employed with a duration of 45 s, and precursors with unassigned or charge state 1 were excluded. An underfill ratio of 1 % was applied.

2.8 | Peptide and Protein Identification

Orbitrap raw MS/MS files were converted to mzML format using msConvert from the ProteoWizard tool suite. Spectral data were analyzed using MSGF+ (v2020.03.14) (Kim and Pevzner 2014) and Percolator (v3.5) (Granholm et al. 2014) for target-decoy Percolator scoring. All searches were performed against the bovine protein database Ensembl 104 within a proteomics workflow (https://github.com/lehtiolab/ddamsproteomics, v2.18), which was constructed using the Nextflow workflow tool (v22.10.5). MSGF+ search parameters included a precursor mass tolerance of 10 ppm, fully tryptic peptides, a maximum peptide length of 50 amino acids, and a maximum charge state of 6. Carbamidomethylation of cysteine residues was set as a fixed modification, while oxidation of methionine residues was considered as a variable modification. MS1 feature detection and quantification were conducted using Hardklor (v2.3.0) and Kronik (v2.20). Peptide spectrum matches (PSMs) identified at a 1 % false discovery rate (FDR) were used to assign gene identities. Protein quantification was based on the average intensity of the top three highest-intensity peptides, with peptide MS1 quantification defined by the 'Best Intensity', representing the intensity value at the apex of the feature elution profile. Note that proteins were included in the proteomics dataset based on this top-three peptide quantification approach, without applying an explicit minimum peptide threshold.

Protein FDRs were calculated using the picked-FDR method, with gene symbols as protein groups, and restricted to 1 % FDR (Savitski et al. 2015). For peptide identification, a precursor ion mass tolerance of 10 ppm and product ion mass tolerances of 0.02 Da for HCD-FTMS. The search algorithm accounted for tryptic peptides with a maximum of two missed cleavages, with carbamidomethylation (C) as a fixed modification and oxidation (M) and deamidation (N) as variable modifications.

2.9 | Missing Data Handling Approach

We designed three conditions to handle the missing values (Figure 2):

1) To reduce the risk of overfitting, we excluded any variable (i.e. a given protein) in which more than half of the values were missing.
2) When all values in a variable were missing in one population, it is possible they are not missing randomly but due to protein absence/low expression in one of the EV populations; when this type of MNAR missing values (i.e. the type in which one population is different from another) was significantly enriched in the dataset, we considered this as biologically meaningful, and imputed the lowest possible value of "1" to replace these values in order to retain the biological differences.
3) In all other cases, missing values were handled using a Random Forest (RF) imputation algorithm.

We devised these three filtering and handling approaches with the following rationale. When most of the values for the same variable are missing, it would be difficult to impute with confidence using any strategy, since there are so few datapoints that can be used as a reference. To avoid biased imputation for those variables with a large number of missing values, we designed the first condition to filter them out (Figure 2, "excluded variables"). Here, we worked with six samples in total, so we removed any variables with at least 4 missing values from the data matrix, that is, more than 50 % of missing data for each variable.

For the second condition, we assumed that when one population, that is, FBS-EVs or milk-EVs, contained no missing values, and the other population contained only missing values, the missing values likely represent low value dropouts, that is, MNAR missing values (protein abundance-dependence) (Jin et al., 2021). This assumption is based on a previous suggestion (Jin et al., 2021) that a variable completely missing in one population is more likely to be MNAR type. We assumed that these values have low abundance, which would mean that in the hypothetical situation where no values were missing and all values were plotted along an x-axis based on abundance, these values would fall on the left side of the plot (i.e. have low scores); for this reason, they have been previously referred to as left-censored missing values (Gardner and Freitas 2021). We reasoned that it is important not to exclude these variables as they likely represent proteins exclusively expressed in one population and therefore are highly meaningful. In this case, all missing values, represented by the placeholder "NaN" (not a number), were replaced by the lowest possible count value: "1" (Figure 2, "variables with one population replaced by value "1""). For example, a given variable, where FBS-EVs = (NaN, NaN, NaN) and milk-EVs = (12, 9, 14) would become FBS-EVs = (1, 1, 1) and milk-EVs = (12, 9, 14). However, importantly, three missing values in one population could also occur randomly (as MAR or MCAR) if for example, the two populations were biologically indistinguishable, protein abundance was generally low or there were analytical issues during sample processing. For this reason, we introduced a statistical test where this replacement by "1" was only implemented if there were significantly more variables containing three missing values in one population than expected by chance.

Finally, we used the RF method to impute the missing values in the remaining variables, that is, those containing one, two or three missing values distributed between both populations, and three missing values in one population if the number of these variables is not significantly different from their predicted occurrence by chance (Figure 2, "variables with imputed values"). We assumed that these missing values are of the MAR or MCAR type (protein abundance-independence). The RF method is an iterative, supervised machine learning algorithm that estimates

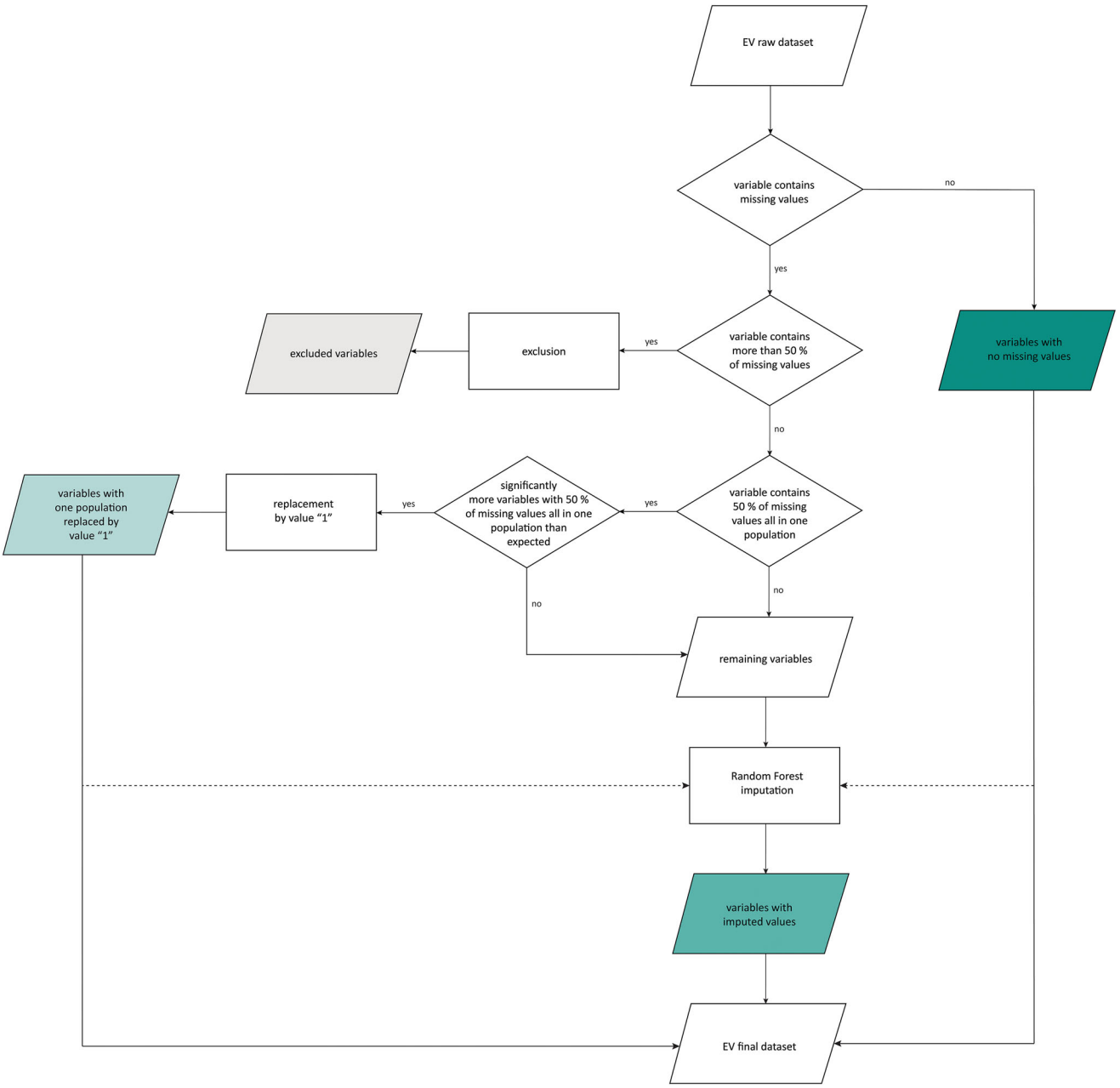


FIGURE 2 | BioProEV workflow.

From the EV raw dataset to the EV final dataset, different categorisations are applied to the variables (i.e. given proteins). According to the number of missing values, the variables are categorised as per the coloured boxes. The dashed arrows indicate that data represented by the associated boxes are included in the dataset used for the Random Forest (RF) imputation, by providing the remaining "forest" of values, but do not themselves undergo imputation by RF. The EV final dataset consists of all data from the green boxes pooled together.

the final imputed values after a sequence of improving predictions (Breiman 2001; Stekhoven and Bühlmann 2012). It is suitable for proteomics data as it has high accuracy in terms of protein abundance and does not require understanding of the missingness mechanisms (Jin et al., 2021). First, it imputes all missing values using the mean calculation. Next, it fits a random forest on the observed data to predict the missing data for each variable with missing values. This process between observed and missing data continues until a stopping criterion or a maximum of iterations is reached, as previously described in detail (Stekhoven and Bühlmann 2012). The variables that had already passed conditions 1 and 2, that is, those originally contained no missing values, and those in which the missing values in one population replaced with value "1" were used for RF imputation. After imputation, the variables and their values passing these three conditions were combined into the "final dataset" (Figure 2, "EV final dataset"). The raw, filtered and final datasets can be found in the supplementary material "S1-fbs-milk-evs-datasets".

To run these handling conditions sequentially, we combined them into a pipeline called BioProEV (***Bio***logically-relevant imputation of missing values in ***Pro***teomic analyses of

Extracellular ***V***esicles) (Miranda et al. 2025). It uses the missForest algorithm (Stekhoven and Bühlmann 2012), which is based on RF imputation technique, and has shown high accuracy handling missing data (Hong and Lynn 2020; Jin et al., 2021; Tang and Ishwaran 2017; Taylor et al. 2022). We included the missForest algorithm in the BioProEV using a Python implementation called MissForest (Yuen 2024), which uses RF features from scikit-learn Python project (Pedregosa et al. 2011, Pedregosa et al. 2024) and has been applied in different types of data (Sacks et al. 2025; Shyam-Sundar et al. 2024). The BioProEV bioinformatics pipeline is implemented in Python language and is freely available at https://doi.org/10.5281/zenodo.15440849, including further details in the "Readme" file.

Please note that this bioinformatics pipeline applies a default thresholding value of 50 % for step 1, that is, any variable in which more than half of the values are missing will be excluded. For purposes of exploration (outlined below in the results section), BioProEV includes an option for the user to adjust this value. The pipeline was developed to analyse the six samples using a threshold of more than 50 % of missing values for exclusion, that is, 4 and 5 NaN. However, without modification, it is also able to handle a different number of samples (e.g. 8, 10); however, it is limited to two populations with the same number of samples and considers only the amount of missing values, not the NaN configuration across samples, e.g. FBS-EVs = (NaN, NaN, NaN) and milk-EVs = (3, 2, NaN) is the same as FBS-EVs = (NaN, 2, NaN) and milk-EVs = (3, NaN, NaN), that is, four missing values in both cases. This means that the pipeline cannot differentiate between NaN configurations, restricting the inclusion/exclusion of specific configurations. Additionally, note that BioProEV was developed to classify MAR/MCAR and MNAR missing values based on patterns in the data, without distinction of sample type, allowing its application to EV datasets from any origin, e.g. serum, milk, blood, urine.

BioProEV was applied to the normalized data based on protein median, where the median protein ratio for each sample is 1. This ensures comparability across samples while preserving relative protein differences. Since imputation depends on the data structure, normalization was performed prior to imputation to avoid the artificial introduction of values that might distort the sample distribution (Karpievitch et al. 2012). Following normalization and replacement/imputation steps, the fold changes between the populations were calculated using the "EV final dataset", and the resulting fold changes were $\log_2$-transformed for downstream analysis.

2.10 | Pathway Enrichment Analysis Using Reactome

We performed pathway enrichment analysis using Reactome, an open-source, manually curated, and peer-reviewed pathway knowledgebase and analysis platform (https://reactome.org/) (Fabregat et al. 2017; Griss et al. 2020; Milacic et al. 2024). Top 100 ranked genes from FBS- and milk-derived EVs identify using LC mass spectrometry were used as input. Analyses were performed with default settings (including species set to Homo sapiens, interactors excluded) and pathways with a p-value ≤ 0.05 were considered significantly enriched. The output included a ranked list of enriched pathways with summary statistics such as found entities and total entities. Voronoi visualization and detailed pathway inspection were conducted using the Reactome Pathway Browser.

2.11 | Graphical Presentation

The data were presented in pie charts and volcano plots using Python (version 3.12). Venn diagrams were first generated in FunRich (http://www.funrich.org/) and presented here using Python (version 3.12). Violin plots were made using GraphPad Prism version 10.1.2 (324). Fisher´s exact test was performed using GraphPad Prism version 10.1.2 (324). To calculate the expected number of proteins within ExoCarta database, the total number of proteins listed in ExoCarta (8877) and the total number of proteins listed in Ensembl 104 database (excluding alternate isoforms) used for protein identification (21880) were used. FBS-EV markers were selected from (X. Zhang et al. 2022) ("List of proteins detected in mouse blood EVs" supplementary), considering only those proteins with "Serum (average)" greater than 0. The selected data were filtered excluding 19 proteins with no gene names (verified in UniProt database [https://www.uniprot.org/]) and 3 duplicated proteins. Milk-EV markers were selected from 38 individual studies gathered in (van Herwijnen et al. 2016). Normality of the data was assessed using the Shapiro-Wilk test for plots in Figure 4B–D. The assumption of normality was met in all cases, and a statistical analysis was conducted using unpaired t-test to evaluate differences between experimental groups, with significance levels denoted as $p < 0.05$ considered significantly different. 307 proteins were selected with "1 NaN: 3 NaN" configuration; however, only 160 proteins showed non-zero variance within populations, allowing them for the statistical analysis. Note that zero variance in both populations for a given protein prevents its assessment of statistical significance. Gene ontology (GO) annotation plots were made with ShinyGO 0.82 (https://bioinformatics.sdstate.edu/go/). Schematics in Figure 1 were made with Biorender.com (agreement number: YH28KPW1Z2) and Reactome Pathway Browser. All final figures were made in Adobe Illustrator (version 29.4. Adobe Creative Cloud). A colour-blind friendly palette was used in all figures.

3 | Results

Three batches of EVs were isolated from FBS and bovine milk respectively, using source-specific isolation methods coupled with SEC (Figure 1A,B). Characterization using NTA revealed that the isolated particles were smaller than 200 nm in diameter, and when coupled to the protein quantification assays, they had reproducible protein: particle ratios (Figure 1C–F). Isolated particles had the expected cup-shaped (collapsed spherical) ultrastructures of EVs visualized by TEM (Figure 1G,H). Proteomic analysis of the six EV samples, with three biological replicates in each of two populations (FBS-EVs and milk-EVs), was performed by LC-ESI-MS/MS. Within the raw dataset, greater than half (85.1 %) of the variables (i.e. given proteins) contained at least one missing value. To handle missing values we devised a three step strategy: (i) we excluded any variable (i.e. a given protein) in which more than half of the values were missing; (ii) if a significantly greater number of variables that only contained

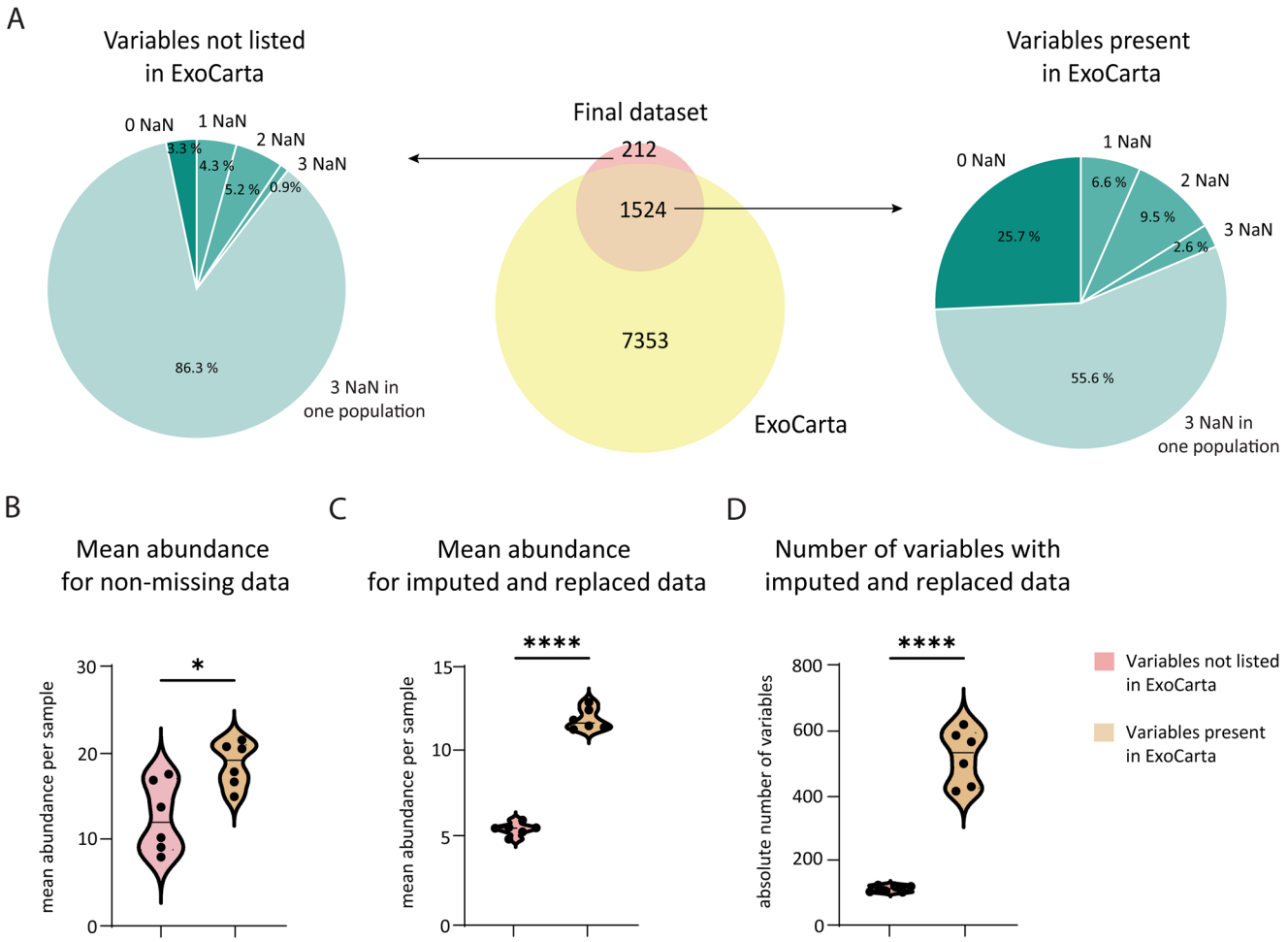


FIGURE 4 | Data distribution in ExoCarta database.

ExoCarta database was used to classify variables in the final dataset, which were analyzed by their original non-missing and missing data rates according to the raw dataset. **A**. Pie charts presenting the distribution of non-missing and missing data for variables classified by their presence in ExoCarta. **B**. Mean protein abundance per sample for variables with only non-missing data; means were statistically compared using unpaired students *t*-test (*p*-value < 0.05). **C**. Mean abundance per sample for variables with imputed and replaced data; means were statistically compared using unpaired students *t*-test (no significant difference). **D**. The absolute number of variables with imputed and replaced data; means were statistically compared using unpaired students *t*-test (*p*-value < 0.0001).

missing values in one population were present than would be expected by chance, missing values were replaced with the lowest possible value of "1"; (iii) in all other cases, missing values were handled using a Random Forest (RF) imputation algorithm (Figure 2). The rationale behind this strategy is outlined in detail in the "Missing data handling approach" section, 2.9 of the Methods. We combined these steps into a bioinformatics pipeline called BioProEV.

Based on the first criterion of BioProEV, 35.1 % of all variables were excluded from the dataset because more than half of the values across the six samples were missing (Figure 3A). All other variables present, the majority of which contained missing values (Figure 3B), were further processed before their eventual inclusion in the final dataset.

BioProEV applies two mechanisms to handle the missing data, depending on where the missing values are located. Three missing values in one population could represent MNAR values due to biological differences in protein expression; if these missing values were MAR or MCAR, then they would have an equal probability to appear in any of the six biological samples (Hong and Lynn 2020). To compare this with the raw dataset, we determined there would be 20 possible configurations (Table S1), where two of those configurations would reflect a situation when all missing values were in one population, giving a probability of 10 % (i.e. 2/20). However, in our raw dataset more than 95 % of the variables containing three missing values were configured in one population, which was significantly different from expected if they had been distributed at random (p <0,0001, using Fisher´s exact test). This analysis indicated that the missing values within the variables containing three missing values in a single population were not missing randomly; we therefore assumed that they were all MNAR values caused by very low protein expression in the sample, and therefore replaced these values with the lowest value in the dataset (i.e. "1") to allow for subsequent quantification (Dabke et al. 2021). Please note that in cases where this assumption would not be met, based on the

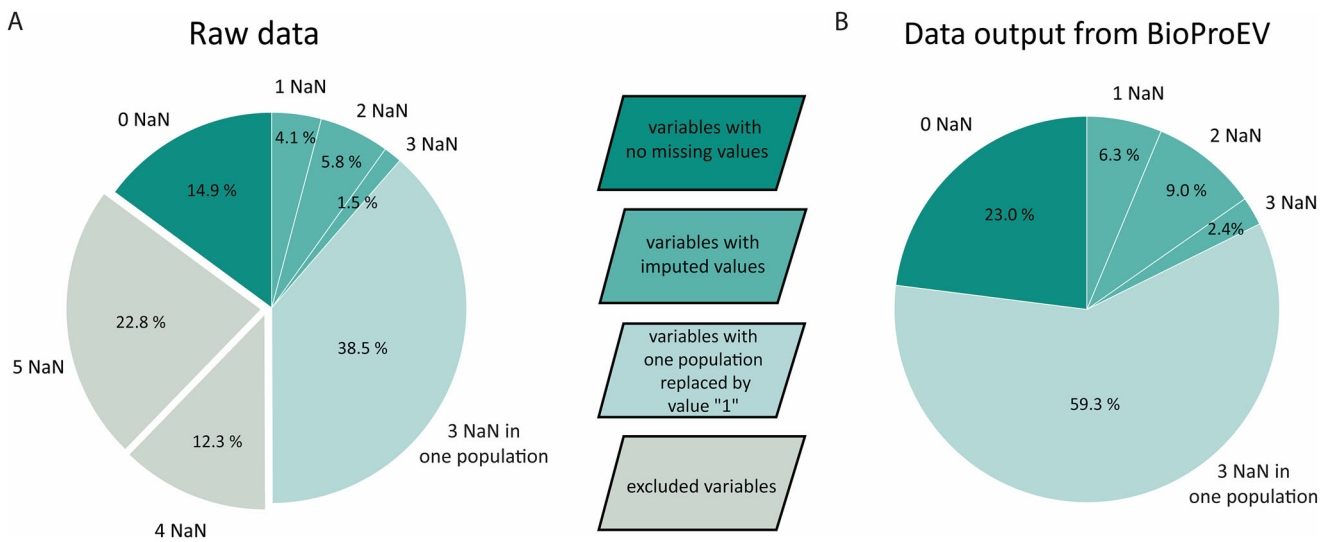


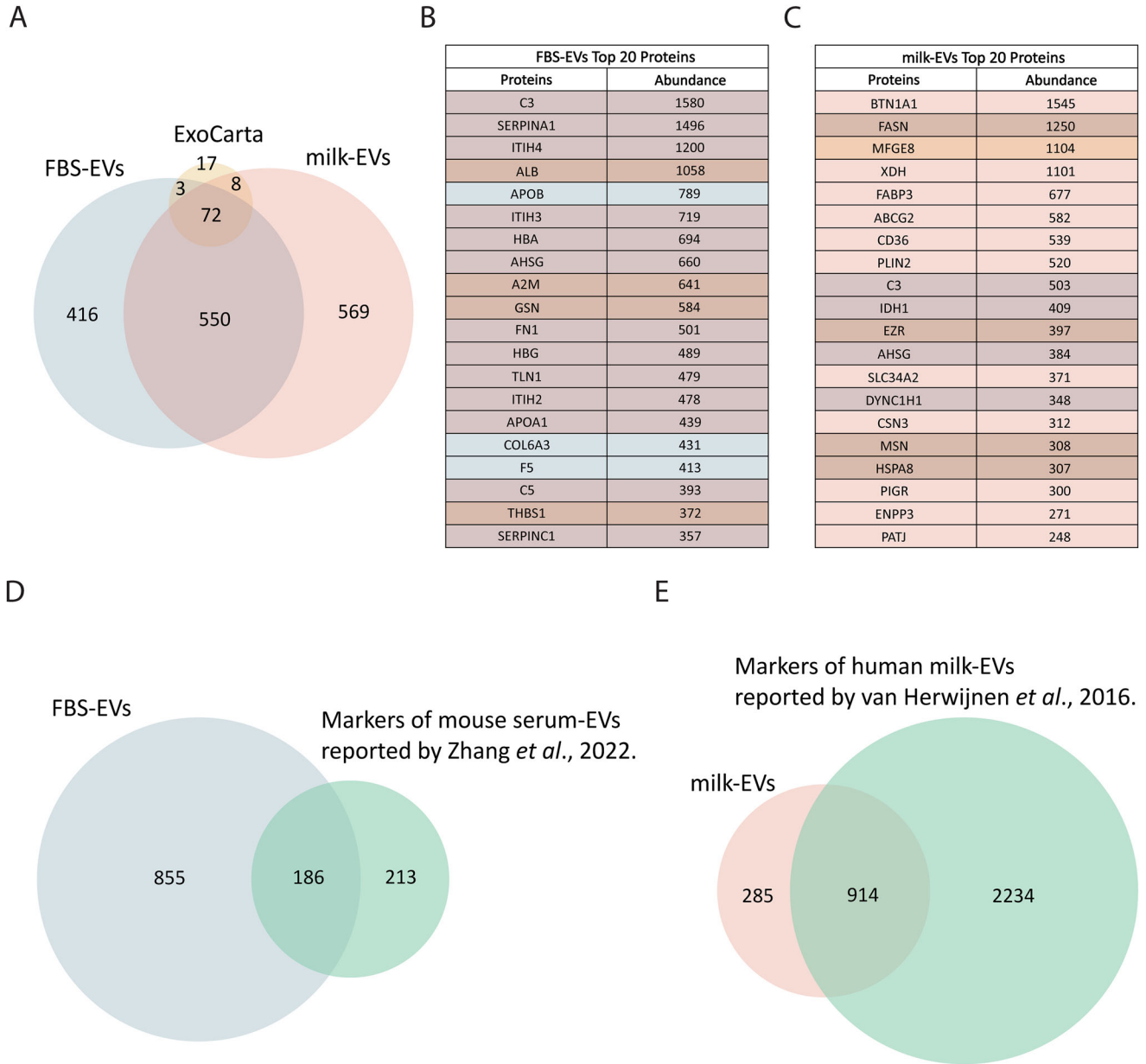
FIGURE 3 | Distribution of non-missing and missing data before and after using BioProEV.
A. Analysis of missing value distribution among variables in raw dataset, showing the variables with observed data, that is, non-missing data (0 NaN), one missing value (1 NaN), two missing values (2 NaN), three missing values distributed in both EV populations (3 NaN), three missing values exclusively in either FBS- or milk-EVs (3 NaN in one population) and those variables with more than half of their values missing (4 NaN and 5 NaN). **B**. The variables with 4 and 5 missing values were excluded from the raw dataset, following the handling by BioProEV, and were therefore excluded in the data output and downstream analyses.

Fisher´s exact test, BioProEV will carry over these missing values and impute them with all remaining missing values using the RF algorithm (Figure 2).

By applying BioProEV to this dataset, we could add 1336 variables to the 400 variables that contained no missing values, which generated a final dataset of 1736 variables. We next explored the composition of proteins sorted by their presence or absence from the ExoCarta database of small EV markers (Keerthikumar et al. 2016; Mathivanan et al. 2012; Mathivanan and Simpson 2009; Simpson et al. 2012). In the final dataset, 87.8 % of proteins were listed in the ExoCarta, indicating that the nanoparticles collected and isolated from FBS- and milk-EVs samples were enriched with expected EV proteins (Figure 4A). Within the final dataset, 25.7 % of the proteins listed in ExoCarta originally contained no missing values in the raw dataset (i.e. NaN = 0), but this number was only 3.3 % for proteins not listed in ExoCarta (Figure 4A). We reasoned that there were two possible reasons for this: either most proteins present in our samples were listed in ExoCarta or the proteins not listed in ExoCarta had a lower abundance and therefore had an increased chance of dropping below detection thresholds and generating MNAR values (Jin et al., 2021). Indeed, among the variables with no missing values, those listed in ExoCarta had a significantly higher abundance than those not listed in this database (Figure 4B). While this was also true for the proteins in which missing values were imputed by BioProEV (Figure 4C), importantly, the majority of proteins (84.7 %) that were included in the final dataset after imputation were listed in ExoCarta (Figure 4D). The number of variables containing imputed values that were listed in ExoCarta was greater than the value that would be expected by chance (84.7 % observed versus 44,4 % expected; p <0,0001, Fisher´s exact test). Together, these results suggest that BioProEV allowed the inclusion of EV-relevant proteins for downstream analyses.

To explore the proteins included in the two populations after missing value handling by BioProEV, we compared them to the 100 most frequently reported small EV markers, according to ExoCarta; a large proportion of proteins (≥ 75%) listed in ExoCarta were present in the FBS- and milk-EVs populations (Figure 5A). Proteins associated with serum (e.g. APOB, F5, ITGA2B) and milk (BTN1A1, MFGE8, XDH) were among the most abundant proteins in each EV preparation (Figure 5B,C). Each population contained expected proteins, based on previous reports of similar EV preparations (van Herwijnen et al. 2016; X. Zhang et al. 2022). Almost half (46.6 %) of serum EV proteins previously described were found in our FBS-EV list (Figure 5D), including the most abundant proteins such as C3, ITIH4 and ALB (Figure 5B). Additionally, more than three quarters (76.2 %) of the milk-EV proteins we identified were reported in previous studies (van Herwijnen et al. 2016) (Figure 5E); the 18 most abundant milk-EV proteins we identified were found within these markers (Figure 5C). These results indicate that in addition to maintaining EV-relevant proteins in the dataset, the final dataset generated by BioProEV maintained source-specific identifiers of EV populations.

The final dataset was used to compare the proteomes of the EV populations. A considerable number of proteins were significantly enriched in either FBS- or milk-EVs, 429 and 621 proteins, respectively. By specifically visualizing the replaced and imputed proteins, the impact of BioProEV became apparent (Figure 6A). Had all variables containing missing values been excluded from the dataset (as is typical in handling missing values by deletion (Han et al. 2023)), only 88 (FBS-EVs) and 161 (milk-EVs) significantly enriched proteins would have been identified (Figure 6A´). However, by using BioProEV, a large number of the proteins only containing values in one of the populations could be included (Fig. 6A´´). Examples include proteins for FBS-EVs



FBS-EVs Top 20 Proteins	
Proteins	Abundance
C3	1580
SERPINA1	1496
ITIH4	1200
ALB	1058
APOB	789
ITIH3	719
HBA	694
AHSG	660
A2M	641
GSN	584
FN1	501
HBG	489
TLN1	479
ITIH2	478
APOA1	439
COL6A3	431
F5	413
C5	393
THBS1	372
SERPINC1	357

milk-EVs Top 20 Proteins	
Proteins	Abundance
BTN1A1	1545
FASN	1250
MFGE8	1104
XDH	1101
FABP3	677
ABCG2	582
CD36	539
PLIN2	520
C3	503
IDH1	409
EZR	397
AHSG	384
SLC34A2	371
DYNC1H1	348
CSN3	312
MSN	308
HSPA8	307
PIGR	300
ENPP3	271
PATJ	248

FIGURE 5 | Most abundant proteins for FBS- and milk-EVs in the final dataset.
A. Venn diagram comparing the most abundant proteins identified in FBS- and milk-EVs, with the 100 most frequently identified proteins in the ExoCarta database. **B** and **C**. The 20 most abundant proteins for both FBS- and milk-EVs, respectively. Common proteins among the top 20 lists for both EV populations are highlighted using the color-code used in A. **D** and **E**. Comparison of proteins identified in FBS- and milk-EVs, with previous reported EV markers (van Herwijnen et al. 2016; Zhang et al. 2022). Please note that for analyses in this Figure, the total number of proteins in each population are not equal because when a protein contained three missing values in one population, it was considered absent and therefore removed from the list of identified proteins in that population.

(e.g. APOB, ITGA2B and F5) and milk-EVs (e.g. BTN1A1, XDH and MFGE8) preserved in the dataset. A further 57 (FBS-EVs) and 75 (milk-EVs) proteins handled via RF imputation were also significantly enriched between the populations (Fig. 6A´´´); these included proteins relevant to the samples, such as C5 and TSG101. Some proteins included by both approaches (replacement and imputation) were found in previously reported FBS- and milk-EV markers, such as ITGA2B (X. Zhang et al. 2022) and CD36 (van Herwijnen et al. 2016). To test the relevance of the differentially enriched proteins, we conducted a gene ontology annotation analysis (Figure 6B; S1); biological processes associated with serum, such as cell adhesion and wound healing, were associated with proteins enriched in the FBS-derived EV preparations, while vesicle-mediated transport and localization were associated with milk-EV preparations, indicating relevant proteins were identified in each population. These annotations were supported by a Reactome analysis of biological pathways (Figure 7).

To further test the validity of the handling approach, we questioned whether the cut-off of 50% NaN used in step 1 might be too

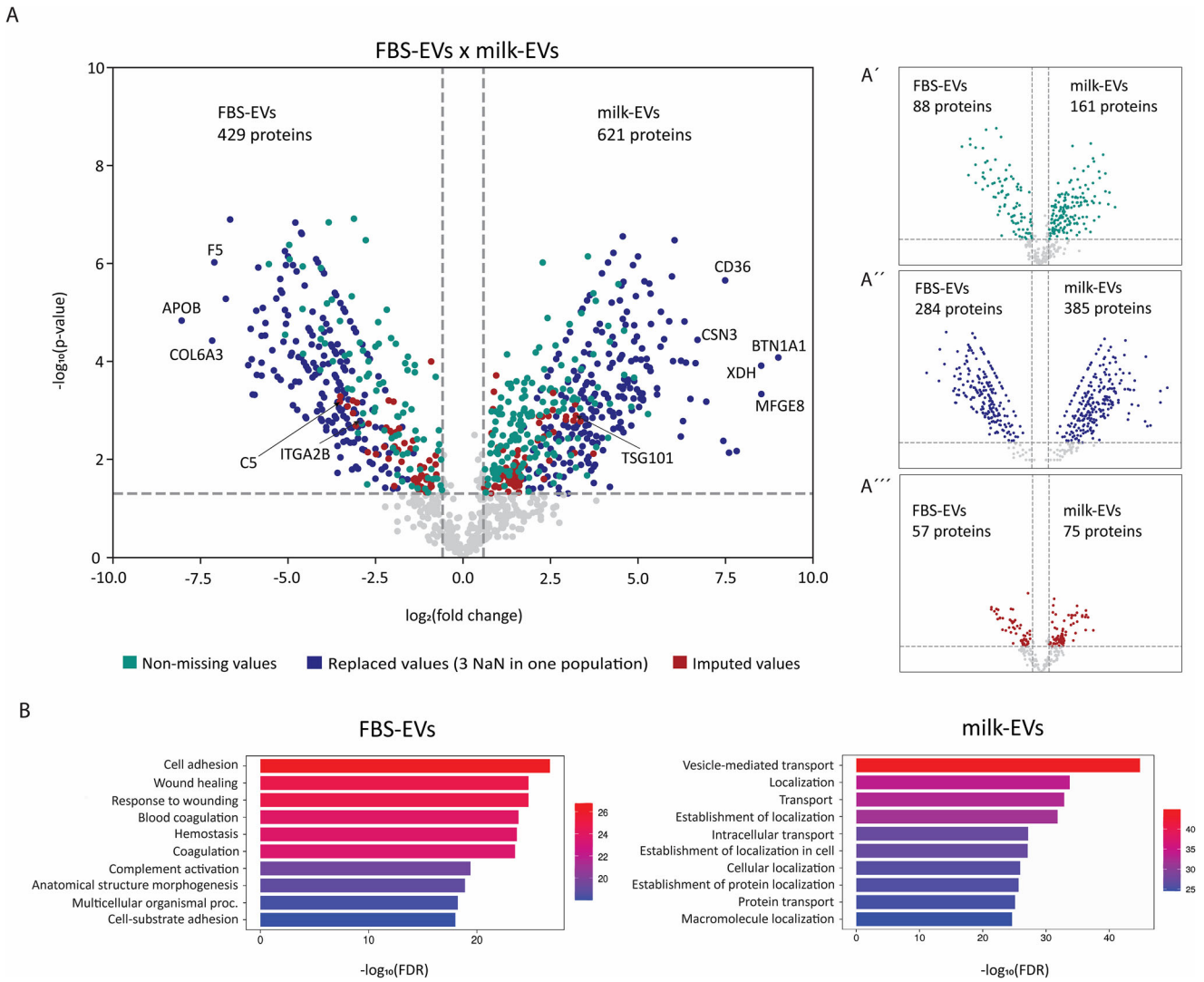


FIGURE 6 | Differentially enriched proteins in FBS- and milk-EV populations.

A. Volcano plot showing the differential expression of proteins in the final dataset. The individual significantly different proteins (spots) are colour-coded based on their handling classification: proteins without missing values (green), proteins with three missing values exclusively in one population (blue), and the proteins with remaining cases of missing values, which were imputed using RF method (red). Gray dots represent non-significant proteins. A´-A´´´ show each classification separately with respective non-significant proteins. **B**. Top 10 biological process annotations for both FBS- and milk-EVs. The plots were made with ShinyGO 0.82 using an FDR cutoff < 0.05.

severe and exclude relevant proteins. To explore this possibility, we isolated the variables with four missing values in a configuration whereby three of the missing values were restricted to one population, e.g. FBS-EVs = (NaN, NaN, NaN) and milk-EVs = (3, 2, NaN), which we refer to as the "1 NaN: 3 NaN" configuration. We chose this configuration based on the prior reasoning that three missing values in one population indicated an absence of the protein (i.e. MNAR), and could be reasonably replaced by "1"; subsequently, the remaining missing value in each variable (MAR/MCAR based on our earlier reasoning) was handled by RF imputation (Fig. S2A). Analysis of the resulting data revealed that 61 proteins were significantly enriched between populations (Fig. S2B); with 32 and 80 of the proteins found in FBS- and milk-EVs listed in ExoCarta, respectively (Fig. S2C). However, it should be noted that in datasets containing a greater proportion of missing values, or when stratification of populations/sub-populations is uncertain, modifying this exclusion threshold could improve the biological relevance of the results. For such analyses or data exploration, simple adjustable thresholding is built in to step 1 of the BioProEV pipeline (detailed in section 2.9, "Missing data handling approach").

Our findings demonstrate that missing value handling is of critical importance to the interpretation of EV proteomics data and that BioProEV, by identifying different types of missing values and handling them accordingly, can be used to maintain proteins of biological relevance in EV proteomics datasets.

FIGURE 7 | Pathway enrichment analysis for FBS- and milk-derived EVs.
A. Pathway enrichment analysis using Reactome for FBS- and milk-derived EVs.

4 | Discussion

Traditional approaches to handling missing values normally include deletion and imputation (Han et al. 2023). The former is a simple case of removing all variables, that is, proteins in the case of proteomics data, with at least one missing value from the dataset. Deletion may introduce severe bias and loss of precision in the downstream analysis, leading to a loss of meaningful information (Little and Rubin, 2019; RUBIN 1976). By contrast, imputation replaces the missing data using either single or multiple imputation strategies, both of which have their pros and cons. The risk of single imputations is that the correlation between variables is ignored, which may lead to an underestimation of the data. For instance, a common single imputation is the mean calculation for each variable, which can be distorted by outliers and reduce the correlation between variables and the variance of the imputed values (Cox and Donnelly 2011; Little and Rubin, 2019). In the case of multiple imputations, the estimated value is predicted from the observed values in an iterative performance, similar to the previous description of RF algorithm. These imputation algorithms use a previous estimated data matrix to run a more accurate prediction (Little and Rubin, 2019). An important point here is that the multiple imputation procedure is able to keep most biologically meaning from data because it considers the relationship between data points. Several methods have been developed, such as RF algorithms (Stekhoven and Bühlmann 2012) and multivariate imputation by chained equations (MICE) (Buuren and Groothuis-Oudshoorn 2011). These methods have

shown a good performance for missing values from MCAR and MAR mechanisms (protein abundance-independence), while MNAR missing values are complex to estimate, since they depend on the missing data (protein abundance-dependence), leading to uncertainty in the results (J. Li et al. 2024). Understanding the missingness mechanism would be ideal to guide how to handle the missing data, however there is no standard rule or test to identify it. Although it is possible to test for MCAR, identifying MAR and MNAR are not testable (Little and Rubin, 2019; van Buuren 2018).

To retain biological information, we took advantage of the possibility to identify MNAR missing values by comparing observed data with expected data using statistical testing. By replacing these MNAR values with the lowest possible value "1" we could retain the biological relevance differences between populations. The advantage of this approach using the example dataset is clear from the color-coded volcano plot in Figure 6, demonstrating the large number of differentially expressed proteins that would have been excluded by a deletion strategy (Hill et al. 2008; Oberg et al. 2008; Wang et al. 2003). While this approach may underestimate true fold changes of a protein between the populations (since the values of absent proteins are likely to be below the value of 1), it has the advantage of increasing the chance that a protein will appear as significantly different between populations during statistical testing (because the variance of the replaced values [i.e. 1, 1, 1] is 0), resulting in meaningful biological differences between the populations. Another advantage of this method was that RF imputation was used for MCAR/MAR but not MNAR missing values, which it can predict with less accuracy (J. Li et al. 2024). Therefore, the two imputation approaches were well suited to the types of missing values they were applied to. Hence, using BioProEV we were able to keep a considerable number of variables in the final dataset for the downstream analysis and maintain biological relevance allowing us to avoid loss of critical information. The imputation of small values is similar to the MinDet method; however, unlike MinDet, we classified missing values by variable, rather than independently, allowing us to better screen those most likely to be MNAR (Gardner and Freitas 2021).

To exemplify the logic of this strategy, if observed values are exclusively present in one population at either high, e.g. protein A: (population #1: 44, 42, 40), (population #2: NaN, NaN, NaN) or moderate levels e.g. protein B: (population #1: 11, 12, 11), (population #2: NaN, NaN, NaN), then all missing values will be replaced by the value "1" (if there are significantly more variables with 50 % of missing values all in one population than would be expected by chance). As the missing values are exclusively in one population (i.e. for both proteins A and B), our assumption is that the missing values are missing not at random (MNAR) and therefore, replacement by "1" will increase the chances of statistical differences between populations, enabling these proteins to be highlighted within the dataset. However, the fold change between protein levels in the populations will be greater for protein A than protein B (i.e. 42 versus 11), retaining the biological information within the observed data in the dataset based on protein abundance. On the other hand, in the theoretical case of protein C (population #1: 11, 8, NaN), (population #2: 10, NaN, NaN), there would be a greater chance that the missing values in the second population would be MAR/MCAR, than those in proteins A and B; therefore, to better retain fold-change differences between populations present in the observed data (in this case 11 and 8 in population #1, and 10 in population #2) random forest imputation was used instead of replacement by the value "1". Even though the number of variables with three missing values in one population occurred more frequently than by chance in the dataset, replacing all of these missing values by "1" introduces a small number of errors into the data. This is because among these variables, some will have three missing values which are MAR or MCAR, and replacing the missing values by "1" may underestimate their true abundance. However, please note that distinguishing MAR and MCAR from MNAR in proteomics data with complete certainty is not possible, and all imputation approaches introduce some discrepancies; we deem this is an acceptable risk in order to retain the true biological differences that this approach allows. To strength the reliability of BioProEV output, independent experimental validation, such as western blotting, enzyme-linked immunosorbent assay (ELISA) or targeted mass spectrometry, could be performed to verify low protein abundance/absence of novel targets appearing exclusively in one EV population.

While there remains no consensus on whether imputation should be performed at the peptide or protein level, in our case, we performed imputation on the protein level; the reason for this choice was because the three top ranked peptide values are combined to generate the protein value, a missing value at protein level indicates multiple peptides with missing values within that protein. This further increases the chance that when a variable contains missing values exclusively in one population at the protein level, these are MNAR (Jin et al., 2021).

The FBS- and milk-derived EV datasets investigated in this work were acquired using LC-ESI-MS/MS in data-dependent acquisition (DDA) mode. This approach selects the most abundant peptides within a range of time to fragmentation and subsequent identification and quantification. This approach tends to fail to detect low-abundance proteins, resulting in a high prevalence of missing values (Zhou et al. 2018). In this case, although missing values still reflect low abundance due to biological reasons, the absence of a protein in the quantified data might be due to a technical factor (MAR type), that is, the limit of detection, and not to a true absence (MNAR type). An alternative approach is the data-independent acquisition (DIA) mode that typically exhibits more complex peptide coverage and lower missing values (Zou et al. 2026). Although DIA mode is increasingly used, less expensive DDA-based techniques are still widely employed, and both approaches have contributed to proteomic discoveries, such as cancer biomarker identification and protein characterization (Kong et al. 2022; Zou et al. 2026). As previously mentioned, the true missingness type cannot be determined, and one population that is entirely absent is more likely to be of the MNAR type than MAR. Therefore, our pipeline design is consistent for DDA data and could, in principle, be extended to DIA data.

Regarding reproducibility, BioProEV relies on the random forest algorithm that contains stochastic elements during training; therefore, it will not always yield precisely the same imputed values when applied to the same dataset. Nevertheless, fold changes and statistical analyses are only likely to be negligibly impacted by these variations.

Pipelines using different types of imputation to differentially handle both MAR/MCAR and MNAR in the same proteomics dataset have previously been called hybrid approaches, among of the most commonly used being KNN-TN (K-nearest neighbours truncation) and MNAR/MAR MI SFI (Multiple Imputation Selection-Filter-Imputation)-hybrid methods. KNN-TN (Shah et al. 2017) assumes that the complete data, that is, no missing values, follows a truncated distribution and considers the detection limit as the truncation point to classify the missingness types. KNN-TN identifies the missingness type by estimating the distribution for the values in each variable; based on this distribution, if missing values in a given variable are predicted to lie at or below the detection limit, they are considered MNAR values, otherwise they are considered to be MAR/MCAR values. Since the missing values are differentiated, all missing values are imputed, in each case variables with similar profiles (i.e. "nearest neighbors") are used to guide imputation. The MNAR/MAR MI SFI-hybrid approach (Gardner and Freitas 2021) is based on the imputeLCMD (Left-Censored Missing Data) R package (Lazar 2014). It classifies MAR/MCAR and MNAR types by assuming that the mean values of variables follow a normal distribution and differences from this are caused by the presence of MNAR values. Then the classified MAR/MCAR values are imputed by KNN method Cover and Hart 1967) and MNAR values by QRILC (quantile regression imputation of left-censored data) method (Lazar 2014), where missing values are imputed separately for each population. While these methods have their strengths, BioProEV has several notable advantages since it (i) excludes variables with large numbers of missing values, (ii) prioritizes the identification of MNAR values based on biological differences between populations, and (iii) uses a highly effective Random Forest machine learning algorithm during imputation. For these reasons we believe BioProEV is a robust pipeline that may be better suited to handle missing values in order to retain biological information within EV proteomics datasets.

5 | Limitations

One reason why there is no standardized approach to missing value handling in proteomics data is that each experimental set-up comes with its own set of assumptions, and a one-size-fits-all strategy cannot be easily justified. Within BioProEV, the user can adjust sample size and the exclusion threshold for the percentage of missing values within a variable (so that variables with as many as 90 % missing values could be retained). Hence, while this means BioProEV can be applied to a variety of datasets, some caution should be observed; for example, in a dataset with two populations of 10 replicates, a threshold of 7/20 observed values (i.e. 65 % missing values) may be justified when 10 of those missing values are in one population and only three are in the other population (since one group contains entirely missing values, it could be biologically relevant to replace those values by "1", with the remaining missing values imputed), but when the missing values are spread between the populations (e.g. 6 missing values in one population, and 7 in the other population), BioProEV will handle these missing values by random forest imputation alone, relying on the few observed data—hence, in this case, confidence in biologically meaningful imputation decreases and the risk of overfitting becomes greater, potentially leading to erroneous results. Furthermore, uncertainty increases if populations contain known or unknown sub-populations. These complex and indeterminate permutations have made us cautious to devise a standard pipeline to handle all missing value configurations. Hence, we have only validated BioProEV up to 50 % missing values, and other researchers must apply their own justification to go beyond this threshold. Nevertheless, the logic and justifications we have devised in this manuscript can be applied to diverse scenarios, depending on the experimental design.

Handling missing values is not a straightforward process, and several aspects must be considered to design an appropriate approach. Here, we addressed the different mechanisms of missingness and how they can be either replaced or imputed, e.g. deletion or mean imputation. BioProEV steps were designed based on prior studies, which demonstrated that the RF method performs well for the MAR/MCAR type and at the protein level, as well as one entirely missing population may be of MNAR type (Jin et al., 2021; Taylor et al. 2022). Furthermore, while BioProEV is expected to be consistent for other datasets, its broader application to DIA datasets has not been evaluated here and could be investigated in future work.

One additional limitation of our approach is that BioProEV applies a statistical test in an attempt to identify MNAR missing value type and, if these missing values occur more frequently than by chance, they are replaced by "1". However, MCAR/MAR can also appear in the same configuration, and this strategy may underestimate the true protein abundance in these variables, introducing bias in the downstream analysis; further development using binomial (Xiao et al. 2013) or detectability-based left-censoring decision (Karpievitch et al. 2009) may improve this thresholding. However, the identification of missingness mechanisms (MCAR/MAR and MNAR) is an existing challenge in handling missing values and the risk in BioProEV workflow is plausible to retain biologically meaningful information.

6 | Conclusion

We highlight the importance of missing value handling to proteomic analysis of EVs, and present a pipeline that can be used to handle such missing values with a focus on retaining biologically-relevant data EVs. We arranged this pipeline, named BioProEV, which contains a built-in decision-making algorithm for hybrid imputation that can be used to deal with EV samples from two biological populations.

Author Contributions

Pâmella Miranda: Conceptualization, Data curation, Investigation, Formal analysis, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review and editing. **Jose G. Marchan-Alvarez**: Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing. **Annemarijn Offens**: Methodology, Writing – review and editing. **Ruihan Zhou**: Methodology, Writing – review and editing. **Mathieu Y. Brunet**: Methodology. **Loes Teeuwen**: Methodology, Writing – review and editing. **Maria Eldh**: Methodology, Writing – review and editing. **Susanne Gabrielsson**: Resources, Conceptualization, Supervision, Writing – review and editing. **Phillip T. Newton**: Conceptualization, Project administration,

Resources, Funding acquisition, Supervision, Writing – original draft, Writing – review and editing.

Acknowledgements

PTN was financially supported by the Novo Nordisk Foundation (0067241), Vetenskapsrådet (2019-01919), Karolinska Institutet, and the European Science Foundation. This project was partially supported by the Fight Kids Cancer Funding Programme, supported by Imagine for Margo, Foundation KickCancer, Foundatioun Kriibskrank Kanner, CRIS Cancer Foundation, and Children Cancer-free Foundation. We thank Dr. Chantal Reinhardt for critical review of the manuscript. Mass spectrometry analysis was performed by the Clinical Proteomics Mass Spectrometry facility, Karolinska Institutet/ Karolinska University Hospital/ Science for Life Laboratory. We also acknowledge the support of the 3D-Electron Microscopy Center at Karolinska Institutet.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Data Availability Statement

BioProEV pipeline can be found at https://doi.org/10.5281/zenodo.15440849, and the FBS- and milk-EVs raw, filtered and final datasets used in this work in the supplementary material "S1-fbs-milk-evs-datasets".

References

Albrecht, D., O. Kniemeyer, A. A. Brakhage, and R. Guthke. 2010. "Missing Values in Gel-Based Proteomics." *Proteomics* 10, no. 6: 1202–1211. https://doi.org/10.1002/pmic.200800576.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32. https://doi.org/10.1023/A:1010933404324.

Buuren, S. v., and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in *R*." *Journal of Statistical Software* 45, no. 3: 1–67. https://doi.org/10.18637/jss.v045.i03.

Chen, C.-L., Y.-F. Lai, P. Tang, et al. 2012. "Comparative and Targeted Proteomic Analyses of Urinary Microparticles From Bladder Cancer and Hernia Patients." *Journal of Proteome Research* 11, no. 12: 5611–5629. https://doi.org/10.1021/pr3008732.

Cover, T., and P. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13, no. 1: 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Cox, D. R. ., and C. A. Donnelly . 2011. *Principles of Applied Statistics*. Cambridge University Press.

Cui, L., D. A. Houston, C. Farquharson, and V. E. MacRae. 2016. "Characterisation of Matrix Vesicles in Skeletal and Soft Tissue Mineralisation." *Bone* 87: 147–158. https://doi.org/10.1016/j.bone.2016.04.007.

Dabke, K., S. Kreimer, M. R. Jones, and S. J. Parker. 2021. "A Simple Optimization Workflow to Enable Precise and Accurate Imputation of Missing Values in Proteomic Data Sets." *Journal of Proteome Research* 20, no. 6: 3214–3229. https://doi.org/10.1021/acs.jproteome.1c00070.

Fabregat, A., K. Sidiropoulos, G. Viteri, et al. 2017. "Reactome Pathway Analysis: A High-Performance in-Memory Approach." *BMC Bioinformatics* 18, no. 1: 142. https://doi.org/10.1186/s12859-017-1559-2.

Gardner, M. L., and M. A. Freitas. 2021. "Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics." *International Journal of Molecular Sciences* 22, no. 17: 9650. https://doi.org/10.3390/ijms22179650.

Granholm, V., S. Kim, J. C. F. Navarro, E. Sjölund, R. D. Smith, and L. Käll. 2014. "Fast and Accurate Database Searches With MS-GF+Percolator." *Journal of Proteome Research* 13, no. 2: 890–897. https://doi.org/10.1021/pr400937n.

Griss, J., G. Viteri, K. Sidiropoulos, V. Nguyen, A. Fabregat, and H. Hermjakob. 2020. "ReactomeGSA—Efficient Multi-Omics Comparative Pathway Analysis." *Molecular & Cellular Proteomics* 19, no. 12: 2115–2125. https://doi.org/10.1074/mcp.TIR120.002155.

Han, J., J. Pei, and H. Tong. 2023. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hill, E. G., J. H. Schwacke, S. Comte-Walters, et al. 2008. "A Statistical Model for iTRAQ Data Analysis." *Journal of Proteome Research* 7, no. 8: 3091–3101. https://doi.org/10.1021/pr070520u.

Hong, S., and H. S. Lynn. 2020. "Accuracy of Random-Forest-Based Imputation of Missing Data in the Presence of Non-Normality, Non-Linearity, and Interaction." *BMC Medical Research Methodology* 20, no. 1: 199. https://doi.org/10.1186/s12874-020-01080-1.

Iraci, N., T. Leonardi, F. Gessler, B. Vega, and S. Pluchino. 2016. "Focus on Extracellular Vesicles: Physiological Role and Signalling Properties of Extracellular Membrane Vesicles." *International Journal of Molecular Sciences* 17, no. 2: 171. https://doi.org/10.3390/ijms17020171.

Jin, L., Y. Bi, C. Hu, et al. 2021. "A Comparative Study of Evaluating Missing Value Imputation Methods in Label-Free Proteomics." *Scientific Reports* 11, no. 1: 1760. https://doi.org/10.1038/s41598-021-81279-4.

Kalluri, R. 2024. "The Biology and Function of Extracellular Vesicles in Immune Response and Immunity." *Immunity* 57, no. 8: 1752–1768. https://doi.org/10.1016/j.immuni.2024.07.009.

Karimi, N., A. Cvjetkovic, S. C. Jang, et al. 2018. "Detailed Analysis of the Plasma Extracellular Vesicle Proteome After Separation From Lipoproteins." *Cellular and Molecular Life Sciences* 75, no. 15: 2873–2886. https://doi.org/10.1007/s00018-018-2773-4.

Karpievitch, Y. V., A. R. Dabney, and R. D. Smith. 2012. "Normalization and Missing Value Imputation for Label-Free LC-MS Analysis." *BMC Bioinformatics* 13, no. S16: S5. https://doi.org/10.1186/1471-2105-13-S16-S5.

Karpievitch, Y., J. Stanley, T. Taverner, et al. 2009. "A Statistical Framework for Protein Quantitation in Bottom-up MS-Based Proteomics." *Bioinformatics* 25, no. 16: 2028–2034. https://doi.org/10.1093/bioinformatics/btp362.

Keerthikumar, S., D. Chisanga, D. Ariyaratne, et al. 2016. "ExoCarta: A Web-Based Compendium of Exosomal Cargo." *Journal of Molecular Biology* 428, no. 4: 688–692. https://doi.org/10.1016/j.jmb.2015.09.019.

Kim, S., and P. A. Pevzner. 2014. "MS-GF+ Makes Progress Towards a Universal Database Search Tool for Proteomics." *Nature Communications* 5, no. 1: 5277. https://doi.org/10.1038/ncomms6277.

Kong, W., H. W. H. Hui, H. Peng, and W. W. B. Goh. 2022. "Dealing With Missing Values in Proteomics Data." *Proteomics* 22: 23–24. https://doi.org/10.1002/pmic.202200092.

Kowal, J., G. Arras, M. Colombo, et al. 2016. "Proteomic Comparison Defines Novel Markers to Characterize Heterogeneous Populations of Extracellular Vesicle Subtypes." *Proceedings of the National Academy of Sciences* 113, no. 8: E968–E977. https://doi.org/10.1073/pnas.1521230113.

Kreimer, S., A. M. Belov, I. Ghiran, S. K. Murthy, D. A. Frank, and A. R. Ivanov. 2015. "Mass-Spectrometry-Based Molecular Characterization of Extracellular Vesicles: Lipidomics and Proteomics." *Journal of Proteome Research* 14, no. 6: 2367–2384. https://doi.org/10.1021/pr501279t.

Kumar, M. A., S. K. Baba, H. Q. Sadida, et al. 2024. "Extracellular Vesicles as Tools and Targets in Therapy for Diseases." *Signal Transduction and Targeted Therapy* 9, no. 1: 27. https://doi.org/10.1038/s41392-024-01735-1.

Lazar, C. 2014. imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation (2.0). R package. https://cran.r-project.org/web/packages/imputeLCMD/imputeLCMD.pdf.

Lazar, C., L. Gatto, M. Ferro, C. Bruley, and T. Burger. 2016. "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies." *Journal of Proteome Research* 15, no. 4: 1116–1125. https://doi.org/10.1021/acs.jproteome.5b00981.

Li, J., S. Guo, R. Ma, et al. 2024. "Comparison of the Effects of Imputation Methods for Missing Data in Predictive Modelling of Cohort Study

Datasets." *BMC Medical Research Methodology* 24, no. 1: 41. https://doi.org/10.1186/s12874-024-02173-x.

Li, W., Y. Li, W. Zhi, et al. 2021. "Diagnostic Value of Using Exosomederived Cysteinerich Protein 61 as Biomarkers for Acute Coronary Syndrome." *Experimental and Therapeutic Medicine* 22, no. 6: 1437. https://doi.org/10.3892/etm.2021.10872.

Lin, S.-Y., C.-H. Chang, H.-C. Wu, et al. 2016. "Proteome Profiling of Urinary Exosomes Identifies Alpha 1-Antitrypsin and H2B1K as Diagnostic and Prognostic Biomarkers for Urothelial Carcinoma." *Scientific Reports* 6, no. 1: 34446. https://doi.org/10.1038/srep34446.

Little, R., and D. Rubin. 2019. *Statistical Analysis With Missing Data*. 3rd ed. Wiley. https://doi.org/10.1002/9781119482260.

Liu, M., and A. Dongre. 2021. "Proper Imputation of Missing Values in Proteomics Datasets for Differential Expression Analysis." *Briefings in Bioinformatics* 22, no. 3: bbaa112. https://doi.org/10.1093/bib/bbaa112.

Marchan-Alvarez, J. G., L. Teeuwen, D. R. Mamand, et al. 2024. "A Protocol to Differentiate the Chondrogenic ATDC5 Cell-Line for the Collection of Chondrocyte-Derived Extracellular Vesicles." *Journal of Extracellular Biology* 3, no. 9: e70004. https://doi.org/10.1002/jex2.70004.

Mathivanan, S., C. J. Fahner, G. E. Reid, and R. J. Simpson. 2012. "ExoCarta 2012: Database of Exosomal Proteins, RNA and Lipids." *Nucleic Acids Research* 40, no. D1: D1241–D1244. https://doi.org/10.1093/nar/gkr828.

Mathivanan, S., and R. J. Simpson. 2009. "ExoCarta: A Compendium of Exosomal Proteins and RNA." *Proteomics* 9, no. 21: 4997–5000. https://doi.org/10.1002/pmic.200900351.

Milacic, M., D. Beavers, P. Conley, et al. 2024. "The Reactome Pathway Knowledgebase 2024." *Nucleic Acids Research* 52, no. D1: D672–D678. https://doi.org/10.1093/nar/gkad1025.

Miranda, P., J. G. Marchan Alvarez, and P. Newton. 2025. BioProEV—Biologically-Relevant Imputation of Missing Values in Proteomic Analyses of Extracellular Vesicles. Zenodo. https://doi.org/10.5281/zenodo.15440849.

Oberg, A. L., D. W. Mahoney, J. E. Eckel-Passow, et al. 2008. "Statistical Analysis of Relative Labeled Mass Spectrometry Data From Complex Samples Using ANOVA." *Journal of Proteome Research* 7, no. 1: 225–233. https://doi.org/10.1021/pr700734f.

Oggero, S., M. de Gaetano, S. Marcone, et al. 2021. "Extracellular Vesicles From Monocyte/Platelet Aggregates Modulate Human Atherosclerotic Plaque Reactivity." *Journal of Extracellular Vesicles* 10, no. 6: 12084. https://doi.org/10.1002/jev2.12084.

Osorio, L. A., M. Lozano, P. Soto, et al. 2024. "Levels of Small Extracellular Vesicles Containing hERG-1 and Hsp47 as Potential Biomarkers for Cardiovascular Diseases." *International Journal of Molecular Sciences* 25, no. 9: 4913. https://doi.org/10.3390/ijms25094913.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion. 2024. Scikit-Learn: Machine Learning in Python. https://scikit-learn.org/stable/.

RUBIN, D. B. 1976. "Inference and Missing Data." *Biometrika* 63, no. 3: 581–592. https://doi.org/10.1093/biomet/63.3.581.

Sacks, D. D., Y. Wang, A. Abron, et al. 2025. "<scp>EEG</scp>Frontal Alpha Asymmetry Mediates the Association Between Maternal and Child Internalizing Symptoms in Childhood." *Journal of Child Psychology and Psychiatry* 66, no. 8: 1129–1140. https://doi.org/10.1111/jcpp.14129.

Savitski, M. M., M. Wilhelm, H. Hahne, B. Kuster, and M. Bantscheff. 2015. "A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets." *Molecular & Cellular Proteomics* 14, no. 9: 2394–2404. https://doi.org/10.1074/mcp.M114.046995.

Shah, J. S., S. N. Rai, A. P. DeFilippis, B. G. Hill, A. Bhatnagar, and G. N. Brock. 2017. "Distribution Based Nearest Neighbor Imputation for Truncated High Dimensional Data With Applications to Pre-Clinical and Clinical Metabolomics Studies." *BMC Bioinformatics [Electronic Resource]* 18, no. 1: 114. https://doi.org/10.1186/s12859-017-1547-6.

Shyam-Sundar, V., G. Slabaugh, S. A. Mohiddin, S. E. Petersen, and N. Aung. 2024. "Clinical Features, Myocardial Injury and Systolic Impairment in Acute Myocarditis." *Open Heart* 11, no. 2: e002901. https://doi.org/10.1136/openhrt-2024-002901.

Simpson, R. J., H. Kalra, and S. Mathivanan. 2012. "ExoCarta as a Resource for Exosomal Research." *Journal of Extracellular Vesicles* 1, no. 1: 18374. https://doi.org/10.3402/jev.v1i0.18374.

Singh, M., P. K. Tiwari, V. Kashyap, and S. Kumar. 2025. "Proteomics of Extracellular Vesicles: Recent Updates, Challenges and Limitations." *Proteomes* 13, no. 1: 12. https://doi.org/10.3390/proteomes13010012.

Stekhoven, D. J., and P. Bühlmann. 2012. "MissForest—Non-parametric Missing Value Imputation for Mixed-type Data." *Bioinformatics* 28, no. 1: 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Tang, F., and H. Ishwaran. 2017. "Random Forest Missing Data Algorithms." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10, no. 6: 363–377. https://doi.org/10.1002/sam.11348.

Taylor, S., M. Ponzini, M. Wilson, and K. Kim. 2022. "Comparison of Imputation and Imputation-free Methods for Statistical Analysis of Mass Spectrometry Data With Missing Data." *Briefings in Bioinformatics* 23, no. 1: bbab353. https://doi.org/10.1093/bib/bbab353.

van Buuren, S. 2018. *Flexible Imputation of Missing Data*. 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259.

Van Deun, J., P. Mestdagh, P. Agostinis, et al. 2017. "EV-TRACK: Transparent Reporting and Centralizing Knowledge in Extracellular Vesicle Research." *Nature Methods* 14, no. 3: 228–232. https://doi.org/10.1038/nmeth.4185.

van Herwijnen, M. J. C., M. I. Zonneveld, S. Goerdayal, et al. 2016. "Comprehensive Proteomic Analysis of Human Milk-Derived Extracellular Vesicles Unveils a Novel Functional Proteome Distinct From Other Milk Components." *Molecular & Cellular Proteomics* 15, no. 11: 3412–3423. https://doi.org/10.1074/mcp.M116.060426.

Wang, W., H. Zhou, H. Lin, et al. 2003. "Quantification of Proteins and Metabolites by Mass Spectrometry Without Isotopic Labeling or Spiked Standards." *Analytical Chemistry* 75, no. 18: 4818–4826. https://doi.org/10.1021/ac026468x.

Webb-Robertson, B.-J. M., H. K. Wiberg, M. M. Matzke, et al. 2015. "Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics." *Journal of Proteome Research* 14, no. 5: 1993–2001. https://doi.org/10.1021/pr501138h.

Welsh, J. A., D. C. I. Goberdhan, L. O'Driscoll, et al. 2024. "Minimal Information for Studies of Extracellular Vesicles (MISEV2023): from Basic to Advanced Approaches." *Journal of Extracellular Vesicles* 13, no. 2: e12404. https://doi.org/10.1002/jev2.12404.

Xiao, C.-L., X.-Z. Chen, Y.-L. Du, X. Sun, G. Zhang, and Q.-Y. He. 2013. "Binomial Probability Distribution Model-Based Protein Identification Algorithm for Tandem Mass Spectrometry Utilizing Peak Intensity Information." *Journal of Proteome Research* 12, no. 1: 328–335. https://doi.org/10.1021/pr300781t.

Yáñez-Mó, M., P. R. M. Siljander, Z. Andreu, et al. 2015. "Biological Properties of Extracellular Vesicles and Their Physiological Functions." *Journal of Extracellular Vesicles* 4, no. 1: 27066. https://doi.org/10.3402/jev.v4.27066.

Yuen, S. Y. H. 2024. MissForest in Python—Arguably the Best Missing Values Imputation Method. Zenodo. https://doi.org/10.5281/zenodo.13368883.

Zhang, H., T. Deng, R. Liu, et al. 2017. "Exosome-Delivered EGFR Regulates Liver Microenvironment to Promote Gastric Cancer Liver

Metastasis." *Nature Communications* 8, no. 1: 15016. https://doi.org/10.1038/ncomms15016.

Zhang, X., T. Takeuchi, A. Takeda, H. Mochizuki, and Y. Nagai. 2022. "Comparison of Serum and Plasma as a Source of Blood Extracellular Vesicles: Increased Levels of Platelet-Derived Particles in Serum Extracellular Vesicle Fractions Alter Content Profiles From Plasma Extracellular Vesicle Fractions." *PLoS ONE* 17, no. 6: e0270634. https://doi.org/10.1371/journal.pone.0270634.

Zhou, L., L. Wong, and W. W. B. Goh. 2018. "Understanding Missing Proteins: A Functional Perspective." *Drug Discovery Today* 23, no. 3: 644–651. https://doi.org/10.1016/j.drudis.2017.11.011.

Zou, X., L. Wang, Y. Chen, et al. 2026. "In-depth Analysis of Data Characteristics and Comparative Evaluation of Dda and Dia Accuracy in Label-Free Quantitative Proteomics of Biological Samples." *Clinical Proteomics* 23, no. 1: 2. https://doi.org/10.1186/s12014-025-09572-2.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supplementary Figure S1: jex270150-sup-0001-FigureS1.tif **Supplementary Figure S2**: jex270150-sup-0002-FigureS2.tif **Supplementary Table S1**: **Table S1**. All possible configurations of observed (Value) and missing value (NaN) across all six EV samples **Supporting Information**: jex270150-sup-0004-SuppMat.xlsx **Supporting Material**: jex270150-sup-0005-SuppMat.docx