Bioinformatics, 2025, **41(12)**, btaf628
https://doi.org/10.1093/bioinformatics/btaf628
Advance Access Publication Date: 20 November 2025
Original Paper

OXFORD

Gene expression

PROTRIDER: protein abundance outlier detection from mass spectrometry-based proteomics data with a conditional autoencoder

Daniela Klaproth-Andrade[1], Ines F. Scheller[1,2], Georgios Tsitsiridis[1], Stefan Loipfinger[1], Christian Mertes[1,3], Dmitrii Smirnov[2,3], Holger Prokisch[2,3,4], Vicente A. Yépez[1], Julien Gagneur[1,2,3,*]

[1]School of Computation, Information and Technology, Technical University of Munich, Garching, 85748, Germany
[2]Computational Health Center, Helmholtz Munich, Neuherberg, 85764, Germany
[3]Institute of Human Genetics, School of Medicine and Health, Technical University of Munich, Munich, 81675, Germany
[4]German Center for Child and Adolescent Health (DZKJ), Partner Site Munich, Munich, 80333, Germany

*Corresponding author. School of Computation, Information and Technology, Technical University of Munich, Garching, 85748, Germany. E-mail: gagneur@in.tum.de
Associate Editor: Macha Nikolski

Abstract

Motivation: Detection of gene regulatory aberrations enhances our ability to interpret the impact of inherited and acquired genetic variation for rare disease diagnostics and tumor characterization. While numerous methods for calling RNA expression outliers from RNA-sequencing data have been proposed, the establishment of protein expression outliers from mass spectrometry data is lacking.

Results: Here, we propose and assess various modeling approaches to call protein expression outliers across three datasets from rare disease diagnostics and oncology. We use as independent evidence the enrichment for outlier calls in matched RNA-seq samples and the enrichment for rare variants likely disrupting protein expression. We show that controlling for hidden confounders and technical covariates, while simultaneously modeling the occurrence of missing values, is largely beneficial and can be achieved using conditional autoencoders. Moreover, we find that the differences between experimental and fitted log-transformed intensities by such models exhibit heavy tails that are poorly captured with the Gaussian distribution and report stronger statistical calibration when instead using the Student's *t*-distribution. Our resulting method, PROTRIDER, outperformed baseline approaches based on raw log-intensities *Z*-scores, PCA, and isolation-based anomaly detection with Isolation forests. The application of PROTRIDER reveals significant enrichments of AlphaMissense pathogenic variants in protein expression outliers. Overall, PROTRIDER provides a method to confidently identify aberrantly expressed proteins applicable to rare disease diagnostics and cancer proteomics.

Availability and implementation: PROTRIDER is freely available at github.com/gagneurlab/PROTRIDER and also available on Zenodo under the DOI zenodo.15569781.

1 Introduction

The detection of outliers in omics data, i.e., values that significantly deviate from the population and can thus be suggestive of a disease-causing gene, is of great importance for rare disease diagnostics (Cummings *et al.* 2017, Kremer *et al.* 2017, Yépez *et al.* 2022, Smail and Montgomery 2024). Importantly, outlier detection in omics data complements genome sequencing data by providing a functional readout to variants of uncertain significance whose interpretation is otherwise inconclusive. Outlier detection methods have been established for RNA-seq abundance, splicing, and chromatin accessibility (Brechtmann *et al.* 2018, Jenkinson *et al.* 2020, Salkovic *et al.* 2020, Mertes *et al.* 2021, Labory *et al.* 2022, Salkovic *et al.* 2023, Scheller *et al.* 2023, Segers *et al.* 2023, Çelik *et al.* 2024). However, DNA accessibility and RNA sequencing cannot capture the effects of all pathogenic variants. Some variants may affect translation or protein stability, without impacting chromatin accessibility or gene expression. To capture those effects, mass spectrometry-based proteomics constitutes an avenue to probe protein abundances as additional functional evidence (Kopajtich *et al.* 2021, Vialle *et al.* 2022, Hock *et al.* 2025, Chui *et al.* 2025). The interest in calling protein expression outliers also extends to cancer research, to characterize alterations in different molecular levels, find biomarkers, and explain drug sensitivities (Roumeliotis *et al.* 2017, Frejno *et al.* 2020).

Several studies have shown that measurements of gene expression, splicing, and chromatin accessibility data exhibit covariation patterns driven by biological and technical factors such as tissue, sampling site within the body, sex, batch, sequencing center, cause of death, sequencer, age, and read length (Kremer *et al.* 2017, Frésard *et al.* 2019, Mertes *et al.* 2021, Yépez *et al.* 2021, Çelik *et al.* 2024). Across those modalities, adjusting for these sources of covariation is strongly beneficial to enrich for the direct regulatory effects of genetic variants. Biological and technical sources of covariation also pertain to labeled proteomics experiments. Notably, samples analyzed together in the same batch of the mass spectrometry run exhibit a stronger correlation than those from different batches, especially for tandem mass tag labeled quantitative proteomics (Brenes *et al.* 2019, Zecha *et al.* 2019, Phua *et al.* 2022). In a previous study, we proposed calling protein level

Received: 3 February 2025; **Revised:** 31 October 2025; **Accepted:** 3 November 2025

outliers using a conditional autoencoder to account for hidden confounders and reported improvements over methods lacking this adjustment (Kopajtich *et al.* 2021).

Here, we expand on and strengthen our previous work and present PROTRIDER. Method-wise, we investigate an alternative strategy to obtain the optimal encoding dimension, model the occurrence of missing values, compare linear against non-linear autoencoders, and perform a statistical assessment based on the Student's *t*-distribution against the Gaussian distribution. Furthermore, we expand the benchmark to two other proteomics datasets of tumor cell lines and to enrichment among expression outliers in matched RNA-seq samples. Finally, we investigate the genetic determinants of the detected aberrant protein abundances, revealing that genes exhibiting protein expression outliers are strongly enriched for missense variants predicted to be pathogenic by AlphaMissense (Cheng *et al.* 2023).

2 Materials and methods

2.1 Datasets and data processing

2.1.1 Mitochondrial disorder dataset

We used a dataset of 143 tandem mass tag (TMT) labeled quantitative proteomics samples with matched RNA-seq samples and variant calls from whole exome sequencing of individuals affected with a rare mitochondrial disorder of suspected genetic origin (Kopajtich *et al.* 2021). This dataset consisted of samples from patient-derived fibroblast cell lines by using a TMT 10-plex labeling reagent. Each TMT batch included 8 patient samples and 2 reference samples. The 143 samples were split over 21 TMT batches, with each batch contributing between 5 and 8 samples, except for one that only contributed one sample. Protein intensities were obtained from protein groups after peptide identification using MaxQuant v.1.6.3.4 (Tyanova *et al.* 2016). The RNA-seq samples were derived from the same fibroblast cultures and reads were counted using DROP (Yépez *et al.* 2021) as previously described (Yépez *et al.* 2022).

2.1.2 Tumor cell line panels

We additionally used proteomics measurements from the two publicly available tumor cell line panels NCI60 ($n = 60$) and CRC65 ($n = 65$), obtained from (Frejno *et al.* 2020). Variant calls from whole exome sequencing from the NCI60 cell lines were downloaded from CellMiner (discover.nci.nih.gov/cellminer/) in the form of the "DNA: Exome Seq—none" processed dataset. For the CRC65 panel, somatic mutation calls from WES were only available for a subset of 33 of the cell lines through the DepMap project (depmap.org/portal/). We downloaded the "OmicsSomaticMutationsProfile.csv" file containing the somatic variant calls, and the files "OmicsProfiles.csv" and "Model.csv" to map from the DepMap profile IDs to the cell line names of the CRC65 data in the proteomics data (Frejno *et al.* 2020). The variant calls for NCI60 were based on the hg19 genome build and annotated with ANNOVAR (Wang *et al.* 2010, Abaan *et al.* 2013), whereas the variants obtained through DepMap were based on the hg38 genome build and annotated with VEP (v. 100.1) among other tools.

2.1.3 Proteomics data preprocessing

TMT-reference samples were excluded in this analysis, and the remaining samples were not normalized using any reference samples. Raw protein intensities were log-transformed and adjusted for overall sample intensity using the DESeq2 size factor normalization (Love *et al.* 2014), resulting in a protein intensity matrix $\mathbf{X}$ with elements $x_{i,j}$ for sample i and protein j.

2.2 Aberrant protein expression level analysis with PROTRIDER

2.2.1 Conditional autoencoder

PROTRIDER uses an autoencoder to capture known and unknown sources of protein intensity variations, yielding expected log-intensities for each protein in each sample. Deviations of the measurements from these expected values are analyzed to identify outliers. First, proteins with more than a defined threshold of missing values across samples were filtered out. That threshold was set to 30% by default, and other thresholds were further investigated. The remaining missing values were set to the protein-wise means in the input matrix $\mathbf{X}$ of the autoencoder and were ignored during mean squared error loss computation. We also considered as possible further input of the autoencoder the binary missingness mask $\mathbf{M}$, in which ones indicate non-missing intensity values and zeros indicate missing values. In this setting, the intensity matrix $\mathbf{X}$ and the missingness mask $\mathbf{M}$ were stacked and jointly fed into the autoencoder, thereby modeling both protein intensities and missing value occurrences simultaneously. We also introduced the option to use a conditional autoencoder approach that explicitly uses specified covariates by including them both in the input of the encoder and the decoder.

The dimension q of the autoencoder bottleneck layer, i.e. latent space, was treated as a hyperparameter and optimized separately. We considered different numbers of layers for the encoder and the decoder, ranging between 1 and 3, where ReLU was used as an activation function between layers of the encoder and decoder, respectively. No non-linear activation was included for 1-layer encoders and decoders, effectively having a linear autoencoder of dimension q. In this case, the 1-layer model, possibly including the missingness mask $\mathbf{M}$ as an additional input, was initialized with truncated Singular Value Decomposition after centering the protein intensity matrix $\mathbf{X}$ protein-wise and with bias terms adjusting for protein-wise means.

The model weights were optimized by minimizing a composite loss function consisting of two terms: (1) the mean squared error (MSE) between the predicted $\hat{\mathbf{X}}$ and observed protein intensities $\mathbf{X}$, computed over all observed values, and (2) the binary cross-entropy (BCE) loss between the predicted probabilities of being observed $\hat{\mathbf{M}}$ and the missingness mask $\mathbf{M}$. These two terms were combined as a weighted sum, with a predefined weighting parameter λ controlling the contribution of the missingness prediction, resulting in a final loss L, defined as

$$L = \mathrm{MSE}\ (\mathbf{X},\ \hat{\mathbf{X}})\ +\ \lambda \cdot \mathrm{BCE}\ (\mathbf{M}, \hat{\mathbf{M}}), \tag{1}$$

where

$$\mathrm{MSE}\ (\mathbf{X},\ \hat{\mathbf{X}}) = \frac{1}{\sum_{i,j} m_{i,j}} \cdot \sum_{i,j} m_{i,j} \cdot \left(x_{i,j} - \hat{x}_{i,j}\right)^2 \tag{2}$$

and

$$\mathrm{BCE}(\mathbf{M}, \hat{\mathbf{M}}) = -\frac{1}{N} \cdot \sum_{i,j} [\ m_{i,j} \cdot\ \log(\hat{m}_{i,j})\ +\ (1 - m_{i,j}) \cdot \log(1 - \hat{m}_{i,j})] \tag{3}$$

with N defined as the total number of proteins times the number of samples in the matrix $\mathbf{X}$, $x_{i,j}$ and $\hat{x}_{i,j}$ as the observed

and predicted intensities and $m_{i,j}$ and $\hat{m}_{i,j}$ as the observed and predicted presence values for each sample i and protein j.

The numerical optimization was performed with Adam (Kingma and Ba 2017) for 400 epochs. For 1-layer autoencoders, a small learning rate of 10^{-4} was used, whereas higher learning rates between 10^{-4} and 10^{-3} were used for multi-layer autoencoders.

2.2.2 Tail probability computation

For each sample i and protein j, the extremeness of the observed pre-processed intensity $x_{i,j}$ relative to the predicted intensity $\hat{x}_{i,j}$ modeled by the autoencoder was quantified using two-sided tail probabilities, denoted $p_{i,j}$. To this end, we considered the residuals of the model $e_{i,j}$ defined as $x_{i,j} - \hat{x}_{i,j}$ and computed their protein-wise means m_j and unbiased standard deviations s_j. The tail probabilities were obtained from either the Gaussian or Student's t-distribution. Gaussian tail probabilities were computed according to

$$p_{i,j} = 2 \cdot \min\left\{\Psi\left(\frac{e_{i,j} - m_j}{s_j}\right), 1 - \Psi\left(\frac{e_{i,j} - m_j}{s_j}\right)\right\}, \quad (4)$$

where Ψ denotes the cumulative function of the normal distribution.

Early investigations with fitting Student's t-distributions with protein-specific degrees of freedom yielded poor statistical calibration, probably due to numerical instability of the likelihood function with respect to the degree of freedom. Therefore, tail probabilities based on a Student's t-distribution were robustly computed with a two-pass approach. In the first pass, we estimated the degrees of freedom, location, and scale parameters of the Student's t-distribution using maximum likelihood for each protein. In the second pass, we set the degrees of freedom for all proteins to a common value $\hat{\nu}_0$ defined as the median of the degrees of freedom estimated in the first pass, and we fitted the location and the scale for each protein again. The two-sided tail probabilities, simply referred to as tail probabilities later on, were calculated as

$$p_{i,j} = 2 \cdot \min\left\{F\left(\frac{e_{i,j} - \hat{\mu}_j}{\hat{\tau}_j}, \hat{\nu}_0\right), 1 - F\left(\frac{e_{i,j} - \hat{\mu}_j}{\hat{\tau}_j}, \hat{\nu}_0\right)\right\}, \quad (5)$$

where F denotes the cumulative function of the Student's t-distribution, $\hat{\mu}_j$ the location estimate, and $\hat{\tau}_j$ the scale estimates of protein j.

Tail probabilities were used to rank outlier candidates. No tail probabilities were reported for missing values.

2.2.3 Selection of the optimal encoding dimension

To find the optimal encoding dimension q of the autoencoder, i.e. the dimension of the autoencoder's latent space, we used two strategies: (i) the optimal hard threshold (OHT) method (Gavish and Donoho 2014), which applies to the linear autoencoders without covariates only, and (ii) a grid search over different values of q.

For the latter approach, at most 25 candidate values or up to half the sample size, whichever is smaller, are explored for finding q. These values are logarithmically spaced between 4 and half the sample size. For each candidate value, we fit the autoencoder after injecting the original dataset with artificial outliers generated with a frequency of 1 per 1000 under a simulation scheme described earlier (Brechtmann *et al.* 2018). Specifically, the outlier intensity $x^o{}_{i,j}$ for sample i and protein j was generated by shifting the observed preprocessed intensity $x_{i,j}$ by $z_{i,j}$ times the standard deviation s_j of $x_{i,j}$. The absolute value $z_{i,j}$ was drawn from a log-normal distribution with the mean of the logarithm equal to 3 and the standard deviation of the logarithm equal to 1.6, and with the sign of $z_{i,j}$ either positive or negative, drawn with equal probability:

$$x^o{}_{i,j} = x_{i,j} + z_{i,j} \cdot s_j \quad (6)$$

We selected the candidate value for q that lead to the highest area under the precision-recall curve (AUPRC) of recovering the previously injected outliers when ranking by tail probabilities. After the optimal encoding dimension was determined either with the OHT or the grid search approach, the autoencoder was fitted using the determined value of q on the actual data without any artificially injected outliers.

2.2.4 Implementation

The autoencoder model of PROTRIDER was implemented in Python (v.3.8.13) using PyTorch (v.1.13.1). It is available at https://github.com/gagneurlab/PROTRIDER. The package includes the Python-based autoencoder implementation, calculates tail probabilities, and produces results tables. We also provide an example dataset and usage guidelines. The code is also available on Zenodo under the DOI https://doi.org/10.5281/zenodo.15569781.

2.3 Tail probability adjustment

2.3.1 Definition of the proportion of false positive calls in a negative control dataset

PROTRIDER does not perform hypothesis testing. Nonetheless, akin to the multiple hypothesis testing problem for P-values, a nominal probability cutoff on tail probabilities would lead to a number of calls increasing with the number of proteins, even in the absence of genuine outliers. We assessed approaches to address this issue by analyzing results on a negative control dataset, in which the observed values were simulated according to the modeling assumptions of PROTRIDER: protein-specific locations linearly related to a latent space, protein-specific scales, and a common value for the degrees of freedom. Any positive call (outlier call) from data generated from the negative control dataset is a false positive. We considered the proportion of false positives among the positive calls and defined this proportion to be equal to 0 if no positive call is made.

2.3.2 Outlier calling using adjusted tail probabilities

While the procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) and the one of Benjamini and Yekutieli (Benjamini and Yekutieli 2001) have theoretical guarantees in the context of multiple hypothesis testing, this does not imply guarantees for our application setting. Therefore, we resorted to empirically assessing whether applications of these procedures to PROTRIDER tail probabilities led to false discovery controls for the negative control dataset we simulated. Specifically, we applied the procedure of Benjamini and Yekutieli, and alternatively, the one of Benjamini and Hochberg, sample-wise, providing tail probabilities instead of P-values, which were originally considered as input in the original publications.

For each sample separately, we considered the unadjusted tail probabilities over the m proteins $p_1, \ldots, p_m$ and performed the following steps:

1) We sorted the tail probabilities in increasing order: $p_{(1)} \leq \ldots \leq p_{(j)} \leq \ldots p_{(m)}$.
2) We computed adjusted tail probabilities: $q_{(j)} = \min_{k \geq j} \frac{m}{k} \cdot c(m) \cdot p_{(j)}$, where $c(m) = 1$ for the Benjamini-Hochberg procedure and $c(m) = \sum_{l=1}^{m} \frac{1}{l}$ for the Benjamini-Yekutieli procedure.
3) We truncated the adjusted tail probabilities at 1, i.e. $q_{(j)} = \min\left(1,\ q_{(j)}\right)$.
4) Finally, we reordered $q_{(j)}$ to match the original order.

This procedure was implemented based on SciPy's false_discovery_control method. Protein outliers were defined as those with an adjusted tail probability of 0.1 or lower.

2.3.3 Empirical assessment of expected proportions of false positive calls in the absence of outliers

We assessed whether the Benjamini-Yekutieli (Benjamini and Yekutieli 2001) and the Benjamini-Hochberg (Benjamini and Hochberg 1995) procedures applied to the tail probabilities (instead of *P*-values) controlled the expected proportion of false positive calls. We demonstrated this for a negative control dataset simulated without any outliers, following a three-step procedure:

1) We first generated data under a model consistent with the PROTRIDER assumptions. To this end, we sampled residuals $e^*_{i,j}$ for all samples i and proteins j from a Student's t-distribution. To obtain simulated data with realistic parameters, we used the locations, scales, and common degrees of freedom estimated on the mitochondrial disorder dataset from the residuals $e_{i,j}$ as described in the Section 2.2.2. The sampled residuals $e^*_{i,j}$ were added to the PROTRIDER fit $\hat{x}_{i,j}$. We then reversed the original preprocessing transformations to yield a synthetic protein intensity matrix with no true outliers.
2) The entire PROTRIDER fitting procedure was applied to each of 100 simulated negative control datasets. Thereby, all parameters were re-estimated, including the degrees of freedom.
3) Next, the Benjamini and Hochberg procedure and the Benjamini and Yekutieli procedure were applied to the tail probability vectors of each sample from each simulated dataset, and outlier calls were made at a fixed threshold.

Since the datasets were simulated without any outliers, any detected outlier constituted a false positive, yielding proportions of false positive calls in each run equal to 0 (no calls made) or 1 (some calls made). The empirical expected proportion of false positive calls was estimated as the average proportion of false positive calls across the 100 simulated datasets and all samples. Moreover, the total amount of calls was recorded for each simulation and sample combination.

2.4 Cross-validation

We also ran PROTRIDER in a leave-one-out and in a 5-fold cross-validation setting. In the leave-one-out setting, each sample was sequentially held out as a test case while the remaining samples were used for model training and hyperparameter tuning. The optimal encoding dimension was determined on the training set using both grid search and OHT approaches. During autoencoder training, 20% of the training set was set aside as a validation set for early stopping. After training, residuals were obtained for all training samples, and for each protein, a Student's t-distribution was fitted in a two-pass approach as described above. The resulting model was then applied to the held-out test sample to derive tail probabilities. This process was systematically repeated for every sample in the dataset.

In the 5-fold setting, the dataset was split into 5 folds. Each fold was used once as a test set while the remaining folds were used for training. The tail probabilities for each test fold were derived using the same procedure as in the leave-one-out setting.

2.5 Comparison to alternative methods

2.5.1 Comparison to PCA

We considered an approach using principal component analysis (PCA) for dimensionality reduction. PCA was applied to the centered, preprocessed protein intensities $x_{i,j}$, and the optimal number of principal components to retain was determined using the OHT criterion applied to the singular values of the data matrix. The data were projected into the subspace spanned by the selected components and subsequently reconstructed into the original feature space by reversing the PCA transformation. Residuals and tail probabilities were then computed from the reconstructed matrix as described above for PROTRIDER.

2.5.2 Comparison to *Z*-scores

We computed *Z*-scores from the preprocessed protein intensities $x_{i,j}$ by subtracting the mean and dividing by the unbiased standard deviation protein-wise. Tail probabilities were calculated using the normal distribution, and the procedure of Benjamini and Yekutieli (Benjamini and Yekutieli 2001) was applied sample-wise.

2.5.3 Comparison to isolation Forest

As an alternative method, we also called outliers by fitting a protein-specific Isolation forest (Liu *et al.* 2012) model in two configurations: (i) fitted directly on the preprocessed intensities $x_{i,j}$, and (ii) fitted on the residuals $e_{i,j}$ returned by PROTRIDER. Outlier candidates were ranked by the anomaly scores predicted by the Isolation Forest models.

2.5.4 Comparison to limma

We used the differential expression detection method limma (Ritchie *et al.* 2015) to call outliers by testing each sample individually against all other samples. To run limma, we used the preprocessed protein intensities $x_{i,j}$ as the input. For each dataset, we included relevant covariates in limma's design matrix, matching those used as input for PROTRIDER. This resulted in a different limma run for each sample of each dataset. The results were concatenated per dataset into a table consisting of multiple testing-corrected *P*-values, fold-changes, and *Z*-scores for each sample-protein combination.

3 Results

3.1 Overview of PROTRIDER

We developed PROTRIDER, a model to call aberrant protein expression from mass spectrometry-based quantitative proteomics data. PROTRIDER models log-transformed protein intensities adjusted for overall sample intensity using a conditional autoencoder to account for biological and technical sources of variation, which may be known or unknown (Fig. 1).

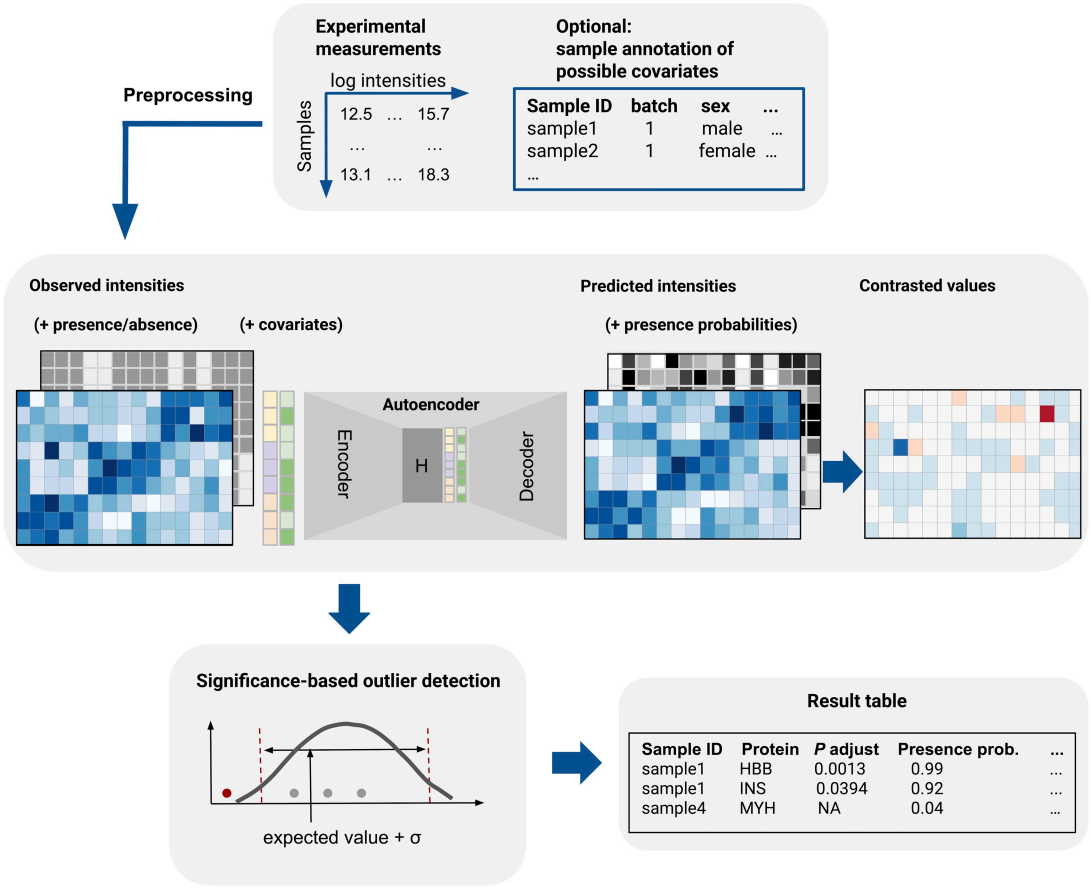

Figure 1. Schematic overview of the PROTRIDER outlier detection approach. PROTRIDER takes a protein intensity matrix from a quantitative proteomics experiment as input, and, optionally, a missingness mask, as well as covariates such as batch and sex, and fits a conditional autoencoder to account for known and unknown biological and technical sources of covariation of proteins across samples. Expected log-transformed protein intensities from the autoencoder are then contrasted with the observed values and tested for statistical significance.

Known covariates, such as batch effects, which are often strongly observed in TMT-labeled proteomics data (Brenes *et al.* 2019), can be directly provided to the model.

Proteomics data often contain missing values, which complicates model fitting (Brenes *et al.* 2019, Phua *et al.* 2022). To alleviate the impact of missing values, we filtered out proteins with more than 30% missing values across an entire dataset. The remaining missing values were ignored in the mean squared error loss computation and tail probability calculations. Moreover, we considered modeling the occurrences of missing values jointly with the observed intensities in the autoencoder (Section 2.2.1).

To identify the optimal latent space dimension, we used two different approaches. The first one conducts a grid search across candidate values for the encoding dimension and selects the one that optimizes the recovery of artificially injected outliers (Section 2.2.3). The second option applies to linear autoencoders without covariates and without missingness modeling only and is based on the Optimal Hard Threshold (OHT) procedure, an analytical solution to the number of principal components to select when assuming the data matrix sums to a low-rank matrix and a white noise matrix (Gavish and Donoho 2014). OHT has been recently used as a computationally efficient alternative to the grid search approach in the context of RNA-seq outlier calling by the OutSingle and saseR methods (Salkovic *et al.* 2023, Segers *et al.* 2023).

After the autoencoder model of PROTRIDER is fitted, the residuals, i.e. the differences between the observations and the autoencoder predictions, are used to compute two-sided tail probabilities assuming either a normal distribution or a Student's *t*-distribution fitted for each protein.

We applied PROTRIDER to three datasets. The first dataset comprises 143 TMT-labelled proteomics samples collected from individuals with a rare, suspected Mendelian, mitochondrial disorder, which we refer to as the mitochondrial disorder dataset (Kopajtich *et al.* 2021). After filtering out proteins with more than 30% missing values across samples, 7060 proteins were quantified (Fig. 1, available as supplementary data at *Bioinformatics* online). Additionally, we considered proteomic measurements of two tumor cell line panels, NCI60 and CRC65 (Frejno *et al.* 2020). The NCI60 panel contains 60 cell lines from various tissues, while the CRC65 panel consists of 65 colorectal tumor cell lines. The two panels were analyzed separately because they differed in the number of proteins detected, tumor entities present in the data, genome build, and exome sequencing processing tool. After removing proteins with more than 30% missing values, the NCI60 dataset was left with 6,755 proteins and the CRC65 dataset with 8,430 proteins (Fig. 2, available as supplementary data at *Bioinformatics* online).

Strong batch effects were observed in the raw log-transformed protein intensities between samples for all three datasets (Fig. 2A, Fig. 3A and C, available as supplementary

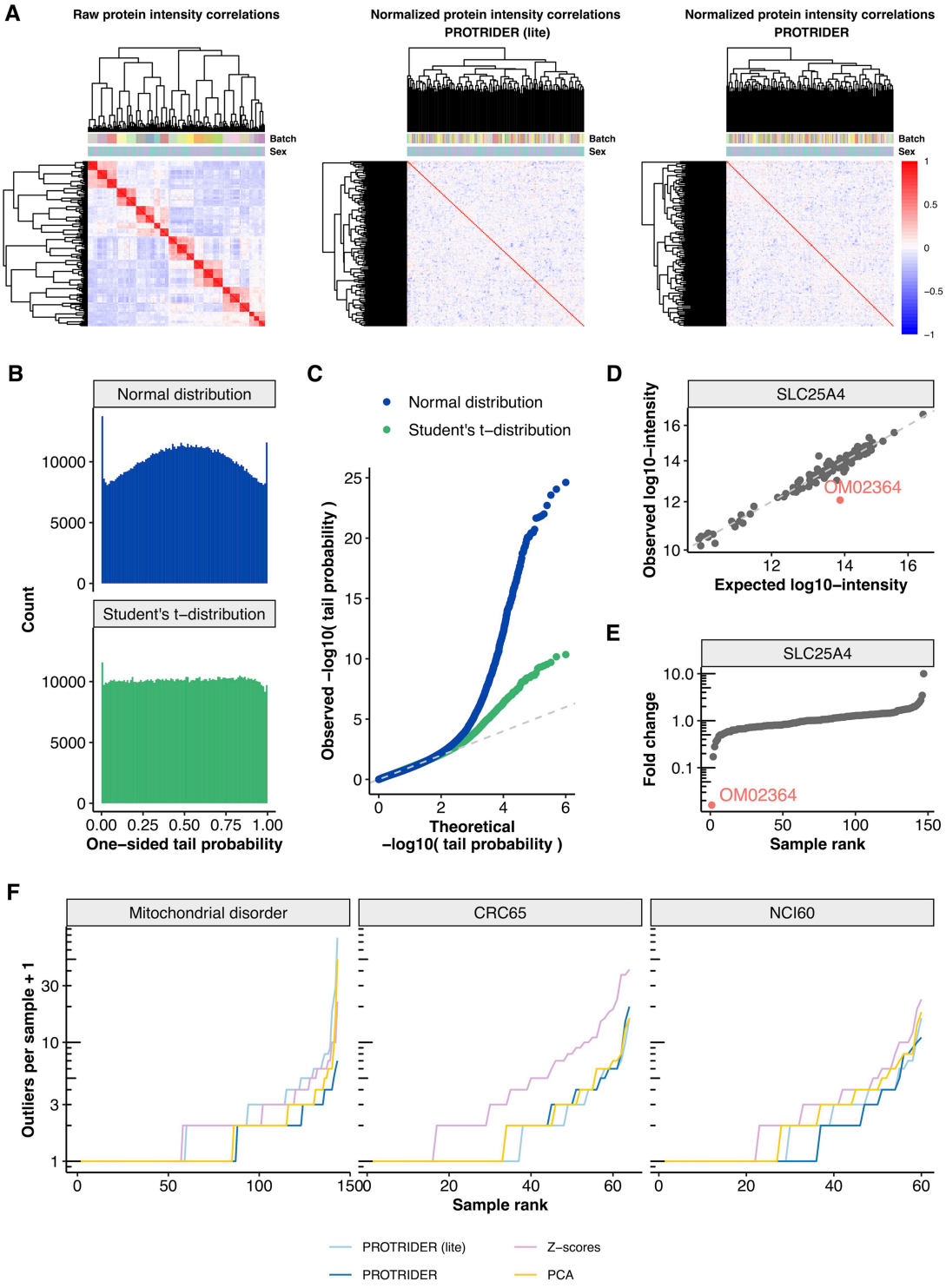

Figure 2. PROTRIDER accounts for known and hidden covariates and typically detects a limited number of outliers per sample. (A) Heatmaps of sample-sample correlations of protein log-transformed intensities on the mitochondrial disorder dataset before (left) and after the PROTRIDER with OHT (middle) and with autoencoder correction using sex and batch as covariates (right). (B) Histogram of tail probabilities from the normal distribution and from the Student's *t*-distribution obtained after the PROTRIDER (lite) correction on the mitochondrial disorder dataset. (C) Quantile-Quantile plot comparing observed $-\log_{10}$(tail probabilities) to their theoretical quantiles under the null (uniform distribution). (D) Observed against expected log-transformed protein intensities for the gene SLC25A4, highlighting the protein underexpression outlier on individual OM02364. (E) Sorted fold changes of protein intensities (ratio of observed and fitted values) for the gene SLC25A4, highlighting individual OM02364. (F) Sorted number of protein abundance outliers per sample obtained by the four methods: PROTRIDER, PROTRIDER (lite), the *Z*-score-based approach, and the approach based on PCA only on the three datasets (facets).

data at *Bioinformatics* online). In the mitochondrial disorder dataset, they were largely driven by the different TMT batches (Fig. 2A). Similarly, in both of the cell line panels, strong correlations between samples were observed in the experimental measurements, driven by the tissue of origin in NCI60 and MSI (Microsatellite Instability Biomarker) status, and subtype classification in CRC65 (Fig. 3A and C, available as supplementary data at *Bioinformatics* online).

To account for such sources of covariations, we evaluated the effect of explicitly fitting a conditional autoencoder model that received the covariates directly as part of its input, in addition to the protein intensities, as well as missingness modeling. We refer to this model as "PROTRIDER". We compared it to an autoencoder that did not have access to information about potential covariates or the missingness mask during model fitting. To this end, we used a linear autoencoder with a dimension determined with the OHT approach. The parameters of this model were further tuned (Section 2.2.3). We named this model "PROTRIDER (lite)." The weights of the autoencoder in both PROTRIDER and PROTRIDER (lite) were initialized using truncated singular value decomposition. On the mitochondrial disorder dataset, we included sex, TMT-batch, sample preparation batch, and sequencing instrument of each sample as covariates, whereas we included sex, age, and tissue of origin for NCI60, and MSI status and subtype classification for CRC65. We found encoding dimensions of 59, 13, and 18 to be optimal for PROTRIDER on the mitochondrial disorder dataset, NCI60, and CRC65, respectively. For PROTRIDER (lite), encoding dimensions of 40, 7, and 5 were obtained, respectively. Notably, we observed that the relationship between the encoding dimension and the performance in recovering artificially injected outliers was typically well-behaved and unimodal, with a relatively wide range of near-optimal values for the encoding dimension (plateau, Section 2.2.3, Fig. 4, available as supplementary data at *Bioinformatics* online). In this context, the area under the precision-recall curve (AUPRC) for recovering artificially injected outliers is used as a relative metric, rather than an absolute measure of performance on real data. We observed a unimodal relationship between AUPRC and encoding dimension, allowing for the reliable selection of the encoding dimension that optimizes recovery on artificial outliers.

Both PROTRIDER versions were able to account for the observed strong batch effects. Specifically, the normalized intensities after fitting the autoencoder models were no longer strongly correlated between samples (Fig. 2A). On the mitochondrial disorder dataset, the median within-batch pairwise sample Spearman correlations were reduced from 0.83 ± 0.059 (mean ± standard deviation, here and elsewhere) to 0.15 ± 0.087 for PROTRIDER (lite) and to 0.13 ± 0.079 for PROTRIDER. These results show that despite not having access to covariates, PROTRIDER (lite) succeeded in capturing the covariation. Similar results were obtained on the tumor cell line panel datasets (Fig. 3B and D, available as supplementary data at *Bioinformatics* online).

When modeling the residuals with protein-specific normal distributions, we observed an excess of one-sided tail probabilities close to 0.5, as well as close to 0 and 1, indicative of the data exhibiting heavy tails (Fig. 2B). Therefore, we opted to use the Student's *t*-distribution instead of the normal distribution. Fitting a Student's *t*-distribution for each protein was robustly achieved by learning a value for the degrees of freedom common to all proteins (Section 2.2.2). The tail probability distributions (observed via histograms and quantile-quantile plots) indicated that the Student's *t*-distribution yielded much better statistical calibration than the normal distribution (Fig. 2B and C).

PROTRIDER does not perform hypothesis testing and therefore does not return *P*-values. Instead, it reports tail probabilities of the model fitted to the data. An important practical goal is to ensure that PROTRIDER rarely makes outlier calls in datasets where no true outliers are present, i.e. to control the proportion of false positives even under a complete null scenario. To empirically assess this, we considered a negative control dataset in which the observed values were generated according to PROTRIDER modeling assumptions, i.e. where the deviations from the autoencoder predictions are drawn independently according to the Student's *t*-distribution. In order to make the simulations realistic, we set the values of the parameters, including the autoencoder predictions, the scales, locations, and common degrees of freedom, to be equal to those we estimated on the mitochondrial disorder dataset. On each simulated dataset, the entire PROTRIDER fitting procedure was applied, i.e. from the latent dimension search to the estimation of the degrees of freedom. Any positive call (outlier call) from the negative control dataset is a false positive. We confirmed that the tail probabilities were approximately uniformly distributed when fitting from scratch PROTRIDER models to 100 such negative control datasets (Section 2.3.3, Fig. 5A and B, available as supplementary data at *Bioinformatics* online). We then applied the Benjamini-Yekutieli (BY, Benjamini and Yekutieli 2001) procedure and the Benjamini-Hochberg (BH, Benjamini and Hochberg 1995) procedure to the tail probabilities, each in a sample-wise manner. Because these procedures were applied to tail probabilities rather than *P*-values, the theoretical False Discovery Rate (FDR) guarantees originally formulated for *P*-values in the hypothesis-testing framework do not apply. Consequently, we empirically evaluated the proportion of false positive calls on the negative control datasets for both the BY and BH procedures. We observed that, when applying the BY procedure with a threshold of 0.1, the average proportion of false positive calls was appropriately controlled and yielded only a small number of false positives (Fig. 5C and D, available as supplementary data at *Bioinformatics* online). In the hypothesis testing framework, the BY procedure is more general than the BH procedure as it can be applied to dependent statistics. In our simulations, the BY procedure behaved more conservatively compared to BH (Fig. 5C, available as supplementary data at *Bioinformatics* online). However, we note that the empirical calibration of the BH procedure might be optimistic because the simulations were performed using independently drawn errors. Therefore, we conservatively used the BY procedure for the subsequent analyses (adjusted tail probability < 0.1).

We exemplify PROTRIDER outlier calls for a rare disease diagnostic relevant case we reported earlier (Kopajtich *et al.* 2021). The mitochondrial solute carrier SLC25A4, which translocates ADP from the cytoplasm into the mitochondrial matrix and ATP from the mitochondrial matrix into the cytoplasm, appeared as a downregulated outlier in the individual OM02364 (adjusted tail probability $= 0.05$). This was evident from the deviation of the observed intensity compared to the PROTRIDER expectation by 61-fold (Fig. 2D and E), which was aberrant compared to the variations seen across

the other individuals. This aberrantly reduced expression of the mitochondrial disorder-causing protein SLC25A4 validated the functional impact of the heterozygous missense candidate variant from the patient and led to the patient's genetic diagnostics (Kopajtich *et al.* 2021).

To provide baseline comparisons, we considered *Z*-scores computed from the observed log-transformed protein intensities adjusted for overall sample intensity (denoted as "*Z*-scores"). We also compared PROTRIDER to an approach that uses only PCA projection with the number of principal components determined via the OHT approach, without optimizing autoencoder weights.

Among all outlier calls (adjusted tail probability of 0.1 or lower), PROTRIDER and PROTRIDER (lite) typically reported 1 outlier per sample across all three datasets (Fig. 2F). This is in line with the number of gene expression outliers typically reported by OUTRIDER on RNA-seq samples (Yépez *et al.* 2022). The PCA-based approach returned a comparable number of outliers per sample to PROTRIDER, while the *Z*-scores approach reported slightly more outliers per sample (median of 1–2 outliers per sample, Fig. 2F).

3.2 Enrichment for rare variants likely disrupting protein expression

To benchmark alternative methods for protein outlier detection performance using orthogonal data, we first considered enrichments for variants that likely lead to protein abundance outliers. To this end, we selected variants that are rare in the human population with a minor allele frequency (MAF) less than or equal to 0.1% in gnomAD (Karczewski *et al.* 2020) and whose consequence is stop, frameshift, splice-site, or missense according to VEP, likely deleterious according to CADD, or pathogenic according to AlphaMissense (McLaren *et al.* 2016, Rentzsch *et al.* 2019, Cheng *et al.* 2023). Those variants do not necessarily lead to aberrant protein abundances. However, one can expect that more accurate protein abundance outlier callers will lead to higher enrichments for genes carrying those variants. We first compared linear autoencoders from PROTRIDER to the PCA-only approach, as well as the *Z*-scores approach, using the variant categories described above as a ground truth proxy.

PROTRIDER outlier calls performed consistently favorably across all variant categories (Fig. 3A).

Specifically, among the 1000 top-ranked underexpression outliers detected by PROTRIDER, 14% ([11.9, 16.3] 95% confidence interval) of the called outliers harbored at least one rare variant whose consequence was stop, frameshift, splice-site, or missense according to VEP on the mitochondrial disorder dataset. The *Z*-score approach showed substantially lower proportions (never exceeding 5%). Very similar results to those obtained with PROTRIDER (lite) were also observed when using a PCA projection alone with the same tail probability computation strategy, without subsequent autoencoder training (Fig. 3A). Moreover, we also considered a non-parametric approach to call and rank outlier candidates. To this end, we applied isolation-based anomaly detection using Isolation Forests (Liu *et al.* 2012). Since our goal was to identify outlier intensities per protein and sample, Isolation Forest was run on each protein individually. We first applied Isolation Forest directly on the preprocessed protein intensities. This led to essentially no enrichment for variants likely disrupting protein expression (Fig. 6, available as supplementary data at *Bioinformatics* online). We next applied Isolation Forest on the residuals returned by PROTRIDER (Fig. 3A), effectively replacing the computed tail probabilities using the Student's *t*-distribution by the returned anomaly scores for ranking. This second approach outperformed the *Z*-scores approach, confirming the importance of adjusting for hidden sources of covariation with the autoencoder, yet did not improve upon PROTRIDER (Fig. 3A). These results are consistent with the good calibration of PROTRIDER tail probabilities evident from the QQ-plot analysis and further indicate that Student's *t*-distributions reasonably capture protein intensity residuals.

Overall, the superiority of PROTRIDER was observed independently of the rank cutoff and for all three considered variant categories (Fig. 3A). Moreover, we compared the results obtained when filtering the dataset to include proteins with at most 30% missing values to those obtained using alternative thresholds for the maximum allowed proportion of missing values per protein. Across all thresholds, PROTRIDER consistently outperformed the other methods (Fig. 7, available as supplementary data at *Bioinformatics* online). There is a trade-off between protein coverage and model performance, whereby a lower proportion of missing values yielded a higher proportion of underexpression outliers associated with rare variants likely disrupting protein expression, yet at the cost of excluding more proteins from the analysis, potentially omitting valuable outlier candidates (Fig. 7, available as supplementary data at *Bioinformatics* online). As a compromise, we opted for a threshold of at most 30% missing values per protein for all subsequent analyses.

For the two tumor cell line datasets, no significant difference between the methods was found when comparing them at equal protein ranks, and the proportions of rare variants likely disrupting protein expression were globally lower compared to the mitochondrial dataset (Fig. 3B and C). One possible explanation for the weaker proportions is the lower sample size (60 and 65 versus 143). Consistent with this hypothesis, the proportions of variants decreased gradually as we downsampled the mitochondrial dataset, yielding proportions at matched sample sizes that were similar to those observed in the tumor cell line dataset ($n = 60$, Fig. 8, available as supplementary data at *Bioinformatics* online). At these lower sample sizes, the performance advantage of PROTRIDER over the other methods also decreased, though it remained slightly more pronounced than in the tumor datasets (Fig. 8, available as supplementary data at *Bioinformatics* online). Additionally, the lower performance observed in the tumor cell lines compared to the mitochondrial disorder dataset may also reflect greater genetic heterogeneity in tumor-derived cell lines relative to patient fibroblasts. We evaluated whether changing the process of artificially injecting outliers during the selection of the encoding dimension could improve the model. Specifically, changing the outlier injection mean from three to six consistently resulted in substantially higher AUPRC when recovering artificially injected outliers for all encoding dimension candidates (Fig. 9A, available as supplementary data at *Bioinformatics* online). However, even though the optimal dimensions changed for the two panels (from 13 to 8 for NCI60 and from 18 to 10 for CRC65), this did not practically result in a difference in the variant enrichment performance (Fig. 9B and C, available as supplementary data at *Bioinformatics* online).

Nonetheless, considering the significant calls only, PROTRIDER reported fewer outliers than the *Z*-scores approach on the tumor cell line panels with an adjusted tail

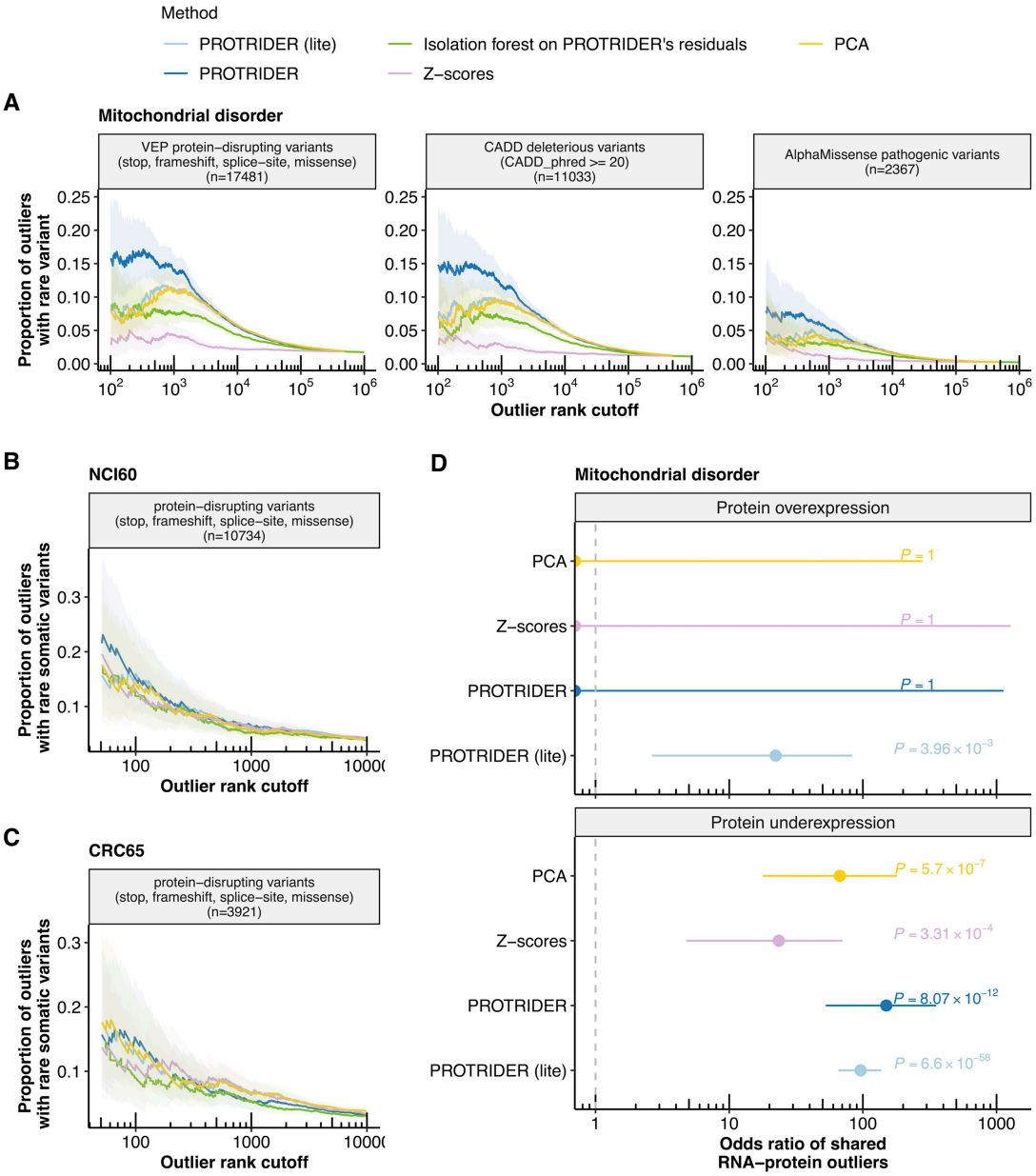

Figure 3. PROTRIDER outperforms baseline approaches on rare variant benchmarks. (A) Proportion of outliers with at least one rare variant on the mitochondrial disorder dataset for underexpression outliers calls from PROTRIDER (blue), PROTRIDER (lite), the isolation forest approach applied on PROTRIDER's residuals, the *Z*-score-based method and a method based on PCA without autoencoder trainingon three sets of rare variant categories as ground truth proxies: (i) VEP stop, frameshift, direct splice-site, and missense variants, (ii) CADD deleterious variants (PHRED score ≥ 20), and (iii) AlphaMissense pathogenic variants. Ribbons mark 95% confidence intervals. (B and C) Same as (A) but for the two tumor cell line panels NCI60 and CRC65, and only on the category of stop, frameshift, direct split-site, and missense variants. (D) Odds ratio of shared RNA and protein outliers (with an adjusted tail probability of 0.1 or lower) in the mitochondrial disorder dataset and their 95% confidence intervals (Fisher's test) for the two PROTRIDER methods, the PCA-based approach, and the *Z*-scores approach.

probability of 0.1 or lower (75 versus 347 on NCI60, 104 versus 157 on CRC65) and maintained, in the case of NCI60, a reasonably high proportion of rare variants likely causing protein abundance aberration compared to the *Z*-scores approach (8.8% versus 7.4% at rank 500, Fig. 3B and C). Altogether, this benchmark using independent evidence from rare genetic variants indicates that PROTRIDER in either setting improves the detection of genuinely aberrantly expressed proteins.

We also considered whether non-linear autoencoders could improve over linear ones. To this end, we added up to three layers with non-linear activations between layers to PROTRIDER. However, using multi-layer models did not improve the performance over the one-layer model on the mitochondrial disorder dataset (Fig. 10, available as supplementary data at *Bioinformatics* online). Therefore, we used one-layer autoencoders for all the subsequent analyses.

Furthermore, we investigated the contribution of various modeling choices by fitting PROTRIDER under different configurations (Fig. 11, available as supplementary data at *Bioinformatics* online). Our analysis showed that modeling missingness as well as using PCA for weight initialization improves PROTRIDER's performance. Interestingly, PROTRIDER (lite)

achieves similar performance regardless of whether PCA-based weight initialization is used. Moreover, we observed consistently similar performance across a range of non-zero weighting factors used to combine the binary cross-entropy with mean squared error loss terms when incorporating the binary missingness modeling (Section 2.2.1), indicating that the model is relatively robust to this parameter choice (Fig. 12, available as supplementary data at *Bioinformatics* online). In contrast, setting the weighting factor to zero, i.e. effectively disabling the missingness modeling, led to a noticeable decrease in performance.

We note that our approach does not require a cross-validation scheme because the encoding dimension is smaller than the dataset dimension, and because the encoding dimension is chosen to optimize the recovery of original, uncorrupted intensities. Nonetheless, we investigated whether a cross-validation scheme could further improve the robustness to PROTRIDER. Specifically, we considered both 5-fold cross-validation and leave-one-out cross-validation (Section 2.4). We observed lower performance for both approaches on the variant enrichment benchmark, whereby the leave-one-out approach performed substantially better than 5-fold cross-validation (Fig. 13, available as supplementary data at *Bioinformatics* online). This consistent trend indicates that the model needs the entire dataset to provide good fits, at least for the sample sizes we could investigate.

3.3 Concordance of protein expression outliers with RNA expression outliers

As another independent type of benchmarking data, we considered gene expression outliers from RNA-sequencing. As for the rare variant benchmark performed above, this benchmark only provides a ground truth proxy, as some gene expression outliers can be translationally buffered. Moreover, some protein abundance outliers may not be reflected in RNA-seq, for instance, due to post-transcriptional regulation (Wang *et al.* 2018, Kusnadi *et al.* 2022). Nevertheless, the proportions of RNA-seq outliers obtained on the same samples can be used to compare different protein abundance outlier callers. For this, we considered outlier calls by PROTRIDER, the PCA-only approach, and the *Z*-scores approach (adjusted tail probability < 0.1). Another added value of using RNA-seq outliers for benchmarking, compared to variants likely disrupting protein expression, is that they allow assessment of the overexpression outliers. All methods obtained a significant enrichment (Fisher's test nominal *P*-value < 0.05) for underexpression RNA-seq outliers called by OUTRIDER (Brechtmann *et al.* 2018). However, the enrichment for the *Z*-scores approach was two times lower than for the other methods, which performed similarly (Fig. 3D). For overexpression, no enrichment for RNA-seq outliers was found by the *Z*-scores approach, nor by the PCA-only approach, nor by PROTRIDER, including missingness modeling, while the remaining methods performed similarly and significantly higher than random (Fig. 3D).

Finally, we considered an approach based on the differential expression test limma (Ritchie *et al.* 2015), applied by testing each sample against all others. We note that, limma is fundamentally designed for group-level hypothesis testing, specifically for comparing mean expression levels between predefined groups, rather than identifying individual outlying samples. Its underlying statistical model targets differences in group means and assumes sufficient replicates per condition, which makes it conceptually distinct from outlier detection methods. Nonetheless, compared to PCA-only or *Z*-score baseline approaches, limma offers the advantage of covariate adjustment. In our benchmarks, applying limma with covariates yielded a strong baseline that outperformed both limma without covariates and the *Z*-scores approach, highlighting the value of adjusting for confounders. Nevertheless, it did not surpass PROTRIDER nor the Isolation forest approach on PROTRIDER's residuals (Figs 7, 8, 14, and 15, available as supplementary data at *Bioinformatics* online).

Taken together, these results show that PROTRIDER is well-suited for the task of aberrant protein abundance detection, outperforming the *Z*-scores approach, the PCA-only approach, the Isolation forest approach, as well as limma with and without covariates on the benchmark using rare protein-disrupting variants and outperforming the limma approach without covariates in the enrichment of shared protein and RNA expression outliers.

3.4 Genetic determinants underlying aberrant protein expression

Having established a protein expression outlier caller, we next investigated various characteristics of the genetic determinants of these outliers. To this end, we considered all rare variants (gnomAD MAF $<0.1\%$) falling within the gene boundary. We note that, as the variants of this dataset were called from whole-exome sequencing, they were biased toward the coding sequence. We found that 30% of the underexpression protein abundance outliers called by PROTRIDER had at least one such rare variant (Fig. 4A). As expected, stop and frameshift variants showed the strongest enrichments (Fig. 4B) and explained overall 15.3% of the underexpression outliers (Fig. 4A). PROTRIDER achieved stronger enrichments, whereas PROTRIDER (lite) showed similar overall patterns, albeit with somewhat reduced signal strength (Fig. 4A and B). No significant enrichments for these variants were found, as expected, for overexpression outliers (Fig. 16, available as supplementary data at *Bioinformatics* online).

We further found a strong enrichment for AlphaMissense pathogenic variants (Fig. 4A and B), which explained another 3.4% of the underexpression outliers. The AlphaMissense categories helped prioritize missense variants affecting protein expression, as the AlphaMissense pathogenic category (odds ratio $= 15.4$) had stronger enrichment than the ambiguous (odds ratio $= 4.2$) and benign (odds ratio $= 0.5$) categories (Fig. 4B).

We evaluated the effect of replacing the conservative Benjamini-Yekutieli method with the Benjamini-Hochberg procedure to adjust tail probabilities. This substitution increased the number of outliers detected by about 69% (208 versus 349) for PROTRIDER and about 134% (360 versus 843) for PROTRIDER (lite), accompanied by only a slight decrease in the enrichment of protein-disrupting rare variants (Fig. 17, available as supplementary data at *Bioinformatics* online).

4 Discussion

We described PROTRIDER, a model that extends the autoencoder-based outlier detection approach already valuable in the RNA-seq-based diagnosis of rare disease patients to mass spectrometry-based proteomics measurements. Using rare genomic variants and RNA-seq outliers as orthogonal data for benchmarking, we showed that protein abundance

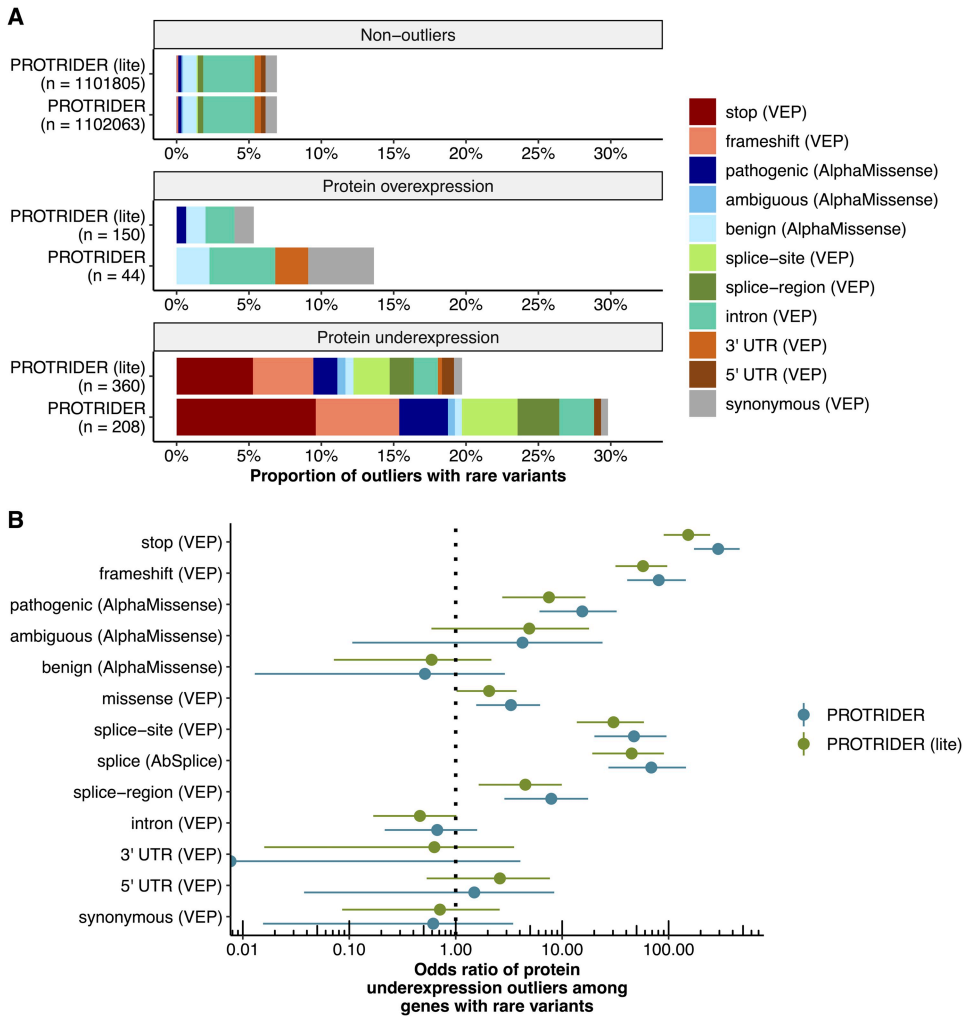

Figure 4. Genetic determinants of protein expression outliers. (A) Proportions of non-outliers, protein overexpression, and protein underexpression outliers with a rare variant detected in the same gene for PROTRIDER and PROTRIDER (lite) in the mitochondrial disorder dataset. Colors indicate the different variant categories based on annotations from VEP and AlphaMissense. (B) Odds ratios and their 95% confidence intervals (Fisher's test) of the enrichment of the proportion for each variant category among underexpression outliers compared to the background proportion of the non-outliers for PROTRIDER and PROTRIDER (lite).

underexpression outliers detected with PROTRIDER outperformed baseline methods based on simple Z-scores, Isolation Forest, and PCA. We found that modeling the occurrence of missing values improved the model performance. In the three investigated datasets from rare disease and tumor cell lines, linear models outperformed non-linear, multi-layer autoencoders. Moreover, we found that the heavy-tailed nature of model residuals was better captured with a Student's t-distribution with a shared degrees-of-freedom parameter across proteins than with a Gaussian distribution. Consequently, using the Student's t-distribution substantially improved the statistical calibration. Finally, we showed that variants predicted by AlphaMissense to be pathogenic, in contrast to the benign predictions, were enriched among underexpression outliers.

The full PROTRIDER model, which uses grid search to optimize the encoding dimension and explicitly incorporates relevant covariates and missingness modeling, consistently achieved the best performance, particularly when benchmarking for rare variants likely disrupting protein expression on the mitochondrial disorder dataset. However, PROTRIDER (lite) remained a robust and computationally efficient alternative. Unlike the full version, PROTRIDER (lite) does not require hyperparameter tuning nor covariate specification, making it a convenient and perhaps more robust starting point. We therefore recommend running both versions. Enrichment for variants likely disrupting protein expression when available, or Q-Q-plots of tail probabilities and sample correlation heatmaps can help decide whether the full model provided an improved fit to the data.

In this study, we presented the main results using the conservative Benjamini-Yekutieli procedure to adjust the tail probabilities, originally introduced to control the false discovery rate in the context of multiple testing under arbitrary dependencies of the statistics. However, the enrichment for variants likely disrupting protein expression on the real data remained high, even though slightly reduced, when using the Benjamini-Hochberg procedure instead of the Benjamini-Yekutieli procedure. Independent of the procedure to adjust tail probabilities, using less stringent cutoffs may be appropriate in scenarios where greater sensitivity to outliers or higher diagnostic rates is a priority. A further plausible scenario in rare disease diagnostics arises when genome analysis yields a candidate variant. Here, the nominal tail probability

of the associated protein may be evaluated directly, without the need for tail probability adjustment.

Given the enrichments we observed for rare and likely deleterious variants of various categories, we expect PROTRIDER to be of value for genetic diagnostic pipelines in addition to RNA-based analyses, especially to capture variant effects acting on the level of translation or protein stability. As we have shown earlier (Kopajtich *et al.* 2021) in the diagnostic context, proteomics can provide functional evidence for rare variants of uncertain significance. This is true for stop and frameshift variants but also for missense variants, which remain difficult variants to interpret. Here, we further showed a strong enrichment of AlphaMissense pathogenic variants in outliers detected at the protein level compared to other missense variants, highlighting a class of variants whose effects are often not captured by RNA-seq analyses.

Our estimation of tail probabilities could, in principle, be improved by separating the data used for model fitting from the data used for evaluating residuals and computing tail probabilities. This separation could enhance the sensitivity of outlier detection by preventing overfitting and ensuring that the model's ability to identify extreme values is not influenced by the same data on which the model was trained. Nevertheless, the empirical distribution of tail probabilities for data simulated under the negative control dataset did not show evidence for overfitting, suggesting the impact of this issue may be limited in practice.

The benchmarks presented in this study focus mostly on underexpression outliers. This is partly due to the nature of the available genomic information, as there were no copy number variations (CNVs) available for the mitochondrial datasets and only for some samples of the CRC65 panel, and partly because it is not straightforward to define classes of variants that lead to an increase in protein abundances. To benchmark overexpression outlier calls with genetic variants, further datasets with available copy number variants and larger sample sizes would be beneficial. Nevertheless, we could show using RNA-seq data that PROTRIDER (lite) overexpression calls were enriched for RNA-seq overexpression calls, indicating that the method can also capture overexpression outliers.

Missing values are common in mass spectrometry-based proteomics, especially in data-dependent acquisition mode (Karpievitch *et al.* 2012, Ahlmann-Eltze and Anders 2019, Brenes *et al.* 2019, Kong *et al.* 2022), and often reflect low-abundance proteins (Kong *et al.* 2022). In PROTRIDER, incorporating the missingness binary mask allowed us to handle missing data more explicitly, rather than simply ignoring it in the loss function. This approach enabled the model to identify biologically relevant absences, such as protein intensities missing in certain samples due to regulation, sample-level, or batch-level technical dropouts. While PROTRIDER models missingness, we did not consider calling individual missing values as outliers. This could be relevant for non-TMT data-dependent acquisition proteomics but would require a substantial extension of the present study. In this context, future work may also investigate other methods, such as variational autoencoders (Collier *et al.* 2020, Nazábal *et al.* 2020) within the scope of outlier detection, particularly for incorporating probabilistic modeling and uncertainty estimates. However, the application of variational autoencoders to sample–protein level outlier detection would require substantial adaptation and evaluation of modeling choices, such as the design of an appropriate conditional prior and regularization terms to preserve reconstruction accuracy.

Furthermore, PROTRIDER only calls outliers at the protein level. However, entire protein complexes are often destabilized (Kremer *et al.* 2017, Kopajtich *et al.* 2021), therefore, additionally assessing outliers at the protein complex level could increase sensitivity. Additionally, peptide-level outliers may offer finer resolution, capturing variant-specific changes such as those caused by post-translational modifications or alternative splicing, which may have important functional consequences. At a more global level, one could also be interested in calling sample-level outliers, i.e. individuals whose entire proteome appears disrupted. In this case, multivariate outlier methods, including LSCP (Locally Selective Combination of Parallel Outlier Ensembles, Berger-Wolf and Chawla 2019) and AnoGAN (Anomaly Detection with Generative Adversarial Networks, Schlegl *et al.* 2017), could be worth investigating. Another interesting extension of our work would be to model longitudinal datasets, which would allow for the detection of outlier values for individual subjects over time.

Although PROTRIDER was developed for proteomics data, it could also be applicable to other mass spectrometry-based measurements, such as metabolomics and lipidomics, which may share similar statistical properties. If the autoencoder-based correction and statistical calibration generalize well, PROTRIDER could support broader applications beyond proteomics.

Acknowledgements

We would like to thank the different users who have used preliminary PROTRIDER versions, as well as the members of the Gagneur lab who have given valuable feedback. We further thank the reviewers for their constructive and insightful comments during the revision process, which significantly helped us improve both the methods and the manuscript.

Author contributions

Daniela Klaproth-Andrade (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [lead], Validation [equal], Visualization [lead], Writing—original draft [equal], Writing—review & editing [equal]), Ines F. Scheller (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [supporting], Validation [equal], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal]), Georgios Tsitsiridis (Formal analysis [supporting], Methodology [equal], Software [supporting], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal]), Stefan Loipfinger (Software [supporting], Writing—review & editing [equal]), Christian Mertes (Data curation [supporting], Writing—review & editing [equal]), Dmitrii Smirnov (Resources [supporting], Writing—review & editing [equal]), Holger Prokisch (Resources [lead], Writing—review & editing [equal]), Vicente A. Yépez (Data curation [equal], Formal analysis [supporting], Software [supporting], Supervision [supporting], Validation [equal], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal]), and Julien Gagneur (Conceptualization [lead], Funding acquisition [lead], Project administration [lead],

Supervision [lead], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: V.A.Y. and C.M. are founders, shareholders, and managing directors of OmicsDiscoveries GmbH. The remaining authors declare that they have no conflict of interest.

Funding

This work was supported by the German Bundesministerium für Bildung und Forschung (BMBF) supported through the VALE (Entdeckung und Vorhersage der Wirkung von genetischen Varianten durch Artifizielle Intelligenz für LEukämie Diagnose und Subtypidentifizierung) project [031L0203B to S.L. and J.G.], through the ERA PerMed project PerMiM [01KU2016B to V.A.Y. and J.G.], through the project CLINSPECT-M [16LW0243K to I.S., D.K.A., J.G.] the European Joint Programme on Rare Diseases, project GENOMIT through the BMBF (01GM2404A to H.P.); EU Horizon 202 project: Recon4IMD [101080997 to H.P.]; by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the project NFDI 1/1 GHGA—German Human Genome-Phenome Archive [#441914366 to C.M. and J.G.] and the IT Infrastructure for Computational Molecular Medicine [#461264291 and #553375143]; by the EVUK program ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria [to I.S. and J.G.]. ERDERA has received funding from the European Union's Horizon Europe research and innovation program under grant agreement N°101156595. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or any other granting authority, who cannot be held responsible for them.

References

Abaan OD, Polley EC, Davis SR *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res* 2013;**73**:4372–82.

Ahlmann-Eltze C, Anders S. proDA: probabilistic dropout analysis for identifying differentially abundant proteins in label-free mass spectrometry. biorxiv, https://doi.org/10.1101/661496, , 2019, preprint: not peer reviewed (1 June 2025, date last accessed).

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J R Stat Soc B* 1995;**57**:289–300.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 2001;**29**:1165–88.

Brechtmann F, Mertes C, Matusevičiūtė A *et al.* OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *Am J Hum Genet* 2018;**103**:907–17.

Brenes A, Hukelmann J, Bensaddek D *et al.* Multibatch TMT reveals false positives, batch effects and missing values. *Mol Cell Proteomics* 2019;**18**:1967–80.

Çelik MH, Gagneur J, Lim RG *et al.* Identifying dysregulated regions in amyotrophic lateral sclerosis through chromatin accessibility outliers. *Hum Genet Genomics Adv* 2024;**5**:100318.

Cheng J, Novati G, Pan J *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492.

Chui MM-C, Kwong AK-Y, Leung HYC *et al.* An outlier approach: advancing diagnosis of neurological diseases through integrating proteomics into multi-omics guided exome reanalysis. *NPJ Genom Med* 2025;**10**:36.

Collier M, Nazabal A, Williams CKI. VAEs in the presence of missing data. arXiv, https://doi.org/10.48550/arXiv.2006.05301, 2020, preprint: not peer reviewed. (1 June 2025, date last accessed).

Cummings BB, Marshall JL, Tukiainen T *et al.*; Genotype-Tissue Expression Consortium. Improving genetic diagnosis in mendelian disease with transcriptome sequencing. *Sci Transl Med* 2017;**9**:eaal5209.

Frejno M, Meng C, Ruprecht B *et al.* Proteome activity landscapes of tumor cell lines determine drug responses. *Nat Commun* 2020;**11**:3639.

Frésard L, Smail C, Ferraro NM *et al.*; Care4Rare Canada Consortium. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* 2019;**25**:911–9.

Gavish M, Donoho DL. The optimal hard threshold for singular values is 4/\sqrt 3. *IEEE Trans Inform Theory* 2014;**60**:5040–53.

Hock DH, Caruana NJ, Semcesen LN *et al.* Untargeted proteomics enables ultra-rapid variant prioritization in mitochondrial and other rare diseases. *Genome Med* 2025;**17**:58.

Jenkinson G, Li YI, Basu S *et al.* LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* 2020;**36**:4609–15.

Karczewski KJ, Francioli LC, Tiao G *et al.*; Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.

Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 2012;**13**:S5.

Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv, https://doi.org/10.48550/arXiv.1412.6980, 2017, preprint: not peer reviewed. (21 January 2025, date last accessed).

Kong W, Hui HWH, Peng H *et al.* Dealing with missing values in proteomics data. *Proteomics* 2022;**22**:e2200092.

Kopajtich R, Smirnov D, Stenton SL *et al.* Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. medRxiv, https://doi.org/10.1101/2021.03.09.21253187, 2021, preprint: not peer reviewed.

Kremer LS, Bader DM, Mertes C *et al.* Genetic diagnosis of mendelian disorders via RNA sequencing. *Nat Commun* 2017;**8**:15824.

Kusnadi EP, Timpone C, Topisirovic I *et al.* Regulation of gene expression via translational buffering. *Biochim Biophys Acta Mol Cell Res* 2022;**1869**:119140.

Labory J, Le Bideau G, Pratella D *et al.* ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data. *Bioinformatics* 2022;**38**:4754–61.

Liu FT, Ting KM, Zhou Z-H. Isolation-based anomaly detection. *ACM Trans Knowl Discov Data* 2012;**6**:1–39.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

McLaren W, Gil L, Hunt SE *et al.* The ensembl variant effect predictor. *Genome Biol* 2016;**17**:122.

Mertes C, Scheller IF, Yépez VA *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun* 2021;**12**:529.

Nazábal A, Olmos PM, Ghahramani Z *et al.* Handling incomplete heterogeneous data using VAEs. *Pattern Recognit* 2020;**107**:107501.

Phua S-X, Lim K-P, Goh WW-B. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput Struct Biotechnol J* 2022;**20**:4369–75.

Rentzsch P, Witten D, Cooper GM *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94.

Ritchie ME, Phipson B, Wu D *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.

Roumeliotis TI, Williams SP, Gonçalves E *et al.* Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell Rep* 2017;**20**:2201–14.

Salkovic E, Abbas MM, Belhaouari SB *et al.* OutPyR: Bayesian inference for RNA-seq outlier detection. *J Comput Sci* 2020; **47**:101245.

Salkovic E, Sadeghi MA, Baggag A *et al.* OutSingle: a novel method of detecting and injecting outliers in RNA-seq count data using the optimal hard threshold for singular values. *Bioinformatics* 2023; **39**:btad142.

Scheller IF, Lutz K, Mertes C *et al.* Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. *Am J Hum Genet* 2023;**110**:2056–67.

Schlegl T, Seeböck P, Waldstein SM *et al.* Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer M, Styner M, Aylward S, *et al.* (eds), *Information Processing in Medical Imaging. Lecture Notes in Computer Science*. Cham: Springer, 2017, 146–57.

Segers A, Gilis J, Van Heetvelde M *et al.* Juggling offsets unlocks RNA-seq tools for fast scalable differential usage, aberrant splicing and expression analyses. biorxiv, https://doi.org/10.1101/2023.06.29.547014, 2023, preprint: not peer reviewed.

Smail C, Montgomery SB. RNA sequencing in disease diagnosis. *Annu Rev Genomics Hum Genet* 2024;**25**:353–67.

Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;**11**:2301–19.

Vialle RA, de Paiva Lopes K, Bennett DA *et al.* Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. *Nat Neurosci* 2022;**25**:504–14.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

Wang SH, Hsiao CJ, Khan Z *et al.* Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol* 2018;**19**:83.

Yépez VA, Gusic M, Kopajtich R *et al.* Clinical implementation of RNA sequencing for mendelian disease diagnostics. *Genome Med* 2022;**14**:38.

Yépez VA, Mertes C, Müller MF *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc* 2021;**16**:1276–96.

Zecha J, Satpathy S, Kanashova T *et al.* TMT labeling for the masses: a robust and cost-efficient, in-solution labeling approach. *Mol Cell Proteomics* 2019;**18**:1468–78.

Zhao Y, Nasrullah Z, Hryniewicki MK *et al.* LSCP: Locally Selective Combination in Parallel Outlier Ensembles. In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. Philadelphia, PA, 2019, 585–93.

Bioinformatics, 2025, 41, 1–14
https://doi.org/10.1093/bioinformatics/btaf628
Original Paper