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

Ines F. Scheller[1,2,*], Daniela Klaproth-Andrade[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] TUM School of Computation, Information and Technology, Technical University of Munich, Garching, Germany

[2] Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany

[3] Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany

[4] German Center for Child and Adolescent Health (DZKJ), partner site Munich, Munich, Germany

* Authors contributed equally.

@ To whom correspondence should be addressed.

Structured 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 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 or on differential expression analysis with limma. 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.

Contact

Julien Gagneur: gagneur at in.tum.de

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-causal gene, is of great importance for rare disease diagnostics (Kremer et al. 2017, Cummings 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 DNA accessibility (Brechtmann et al. 2018, Mertes et al. 2021, Scheller et al. 2023, Çelik et al. 2024, Salkovic et al. 2023, Segers et al. 2023, Jenkinson et al. 2020, Labory et al. 2022, Salkovic et al. 2020). However, DNA accessibility and RNA sequencing cannot capture the effects of all pathogenic variants. Some variants may affect translation or protein stability, without impacting DNA 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. 2024). 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 (Frejno et al. 2020, Roumeliotis et al. 2017).

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 covariations 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, Lim and Goh 2022). In a previous study, we proposed calling protein level 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, 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).

Methods

Datasets

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, Temu and Cox 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).

Tumor cell line panels

We additionally used proteomics measurements of 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 (Abaan et al. 2013, Wang, Li and Hakonarson 2010), whereas the variants obtained through DepMap were based on the hg38 genome build and annotated with VEP (v. 100.1) among other tools (DepMap 24Q2 Public 2024).

Proteomics data preprocessing

Raw protein intensities were log-transformed and adjusted for overall sample intensity using the DESeq2 size factor normalization (Love, Huber and Anders 2014) resulting in a protein intensity matrix $\mathbf{X}$ with elements $x_{i,j}$ for sample i and protein j.

Aberrant protein expression level analysis with PROTRIDER

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 to these expected values are analyzed to call outliers. First, proteins with more than a defined threshold (30%) of missing values across samples were

filtered out. The remaining missing values were set to the geometric mean of the protein intensities in the input of the autoencoder and were ignored during loss computation.

The dimension *q* of the autoencoder bottleneck layer 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. 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.

No non-linear activation was included for 1-layer encoders and decoders, effectively having a linear autoencoder of dimension *q*. In this case, the model 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 weights were numerically optimized using the mean-squared error loss computed over all non-missing values. The numerical optimization was performed with Adam (Kingma and Ba 2017) for 100 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.

<u>P-value computation</u>

For each sample i and protein j, statistical testing of the observed pre-processed intensities $x_{i,j}$ with respect to the fitted intensities $\hat{x}_{i,j}$ modeled by the autoencoder was performed using two-sided *P*-values 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 μ_j and unbiased standard deviations σ_j. The *P*-values were obtained from either Gaussian or Student's t-distribution. Gaussian *P*-values were computed according to

$$p_{i,j} = 2 \cdot \min\left\{\Psi\left(\frac{e_{i,j}-\mu_j}{\sigma_j}\right),\ 1 - \Psi\left(\frac{e_{i,j}-\mu_j}{\sigma_j}\right)\right\}. \tag{1}$$

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, *P*-values 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 degree of freedom for all proteins to a common value ν_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 *P*-values were calculated as

$$p_{i,j} = 2 \cdot \min\left\{F\left(\frac{e_{i,j}-\mu_j}{s_j},\ \nu_0\right),\ 1 - F\left(\frac{e_{i,j}-\mu_j}{s_j},\ \nu_0\right)\right\}, \tag{2}$$

where F denotes the cumulative function of the Student's t-distribution, μ_j the location, and s_j the scale of protein j.

No *P*-values were reported for missing values. False Discovery Rates (FDR) were obtained by applying the method of Benjamini and Yekutieli (Benjamini and Yekutieli 2001) sample-wise. Protein outliers were defined as those with an FDR of 0.1 or lower.

Computation of optimal hidden dimension

To find the optimal encoding dimension *q* of the autoencoder, we employed 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, we evaluated the optimality of each *q* based on the area under the precision-recall curve of recovering artificially injected outliers when ranking by *P*-values. For this purpose, artificial outliers were generated with a frequency of 1 per 1,000 obtained with 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:

$$x^o_{i,j} = x_{i,j} + z_{i,j} \cdot s_j \tag{3}$$

Implementation

The autoencoder model of PROTRIDER was implemented in Python (v.3.8.13) using PyTorch (v.1.13.1). It is available at github.com/gagneurlab/PROTRIDER. The package includes the Python-based autoencoder implementation, calculates *P*-values, and produces results tables. We also provide an example dataset and usage guidelines.

Limma benchmark

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. Relevant covariates for each dataset were included through limma's design matrix. 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.

Z-score benchmark

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. *P*-values were calculated using the normal distribution and multiple-testing correction was performed using Benjamini and Yekutieli's method (Benjamini and Yekutieli 2001).

Variant annotations

To evaluate PROTRIDER on independent benchmarks, we considered different types of rare variant categories. Variants were considered rare if they had a minor allele frequency of 0.1% or lower according to gnomAD v2.1.1, (Karczewski et al. 2020). Variants on the mitochondrial disorder dataset were annotated using the consequences from the Variant Effect Predictor VEP v99 (McLaren et al. 2016), AbSplice (Wagner et al. 2023), AlphaMissense (Cheng et al. 2023), and CADD v1.6 (Rentzsch et al. 2019). For VEP annotations, when more than one rare variant was present in a gene for a given sample, we considered only the most severe consequence based on Ensembl's severity ranking (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html). AlphaMissense classifications were only considered for gene-sample combinations without any variant categorized as high-impact by VEP in the same gene and replaced the VEP missense annotations.

On the cancer cell line datasets, we used the variant annotations containing ANNOVAR consequences for the NCI60 samples and VEP annotations for the CRC65 samples. Additionally, we annotated AlphaMissense predictions in both datasets. As the variant calls corresponding to the cancer cell lines were somatic variant calls, we considered only variants with a sample allele frequency of at least 50%.

Enrichment for rare variants likely disrupting protein abundance

To benchmark underexpression outlier calls, we considered three non-disjoint sets of rare variants that could disrupt protein abundance: i) stop, frameshift, direct splice-site, and missense variants annotated by VEP, ii) likely deleterious variants based on CADD with a CADD PHRED-scaled score of at least 20, and iii) variants predicted as pathogenic by AlphaMissense. For each set, proportions were computed among all underexpression outliers.

Aberrant expression analysis of RNA-seq samples

To compare the protein abundance outlier calls to RNA expression outliers in the mitochondrial disorder dataset, we ran the gene expression outlier detection method OUTRIDER (Brechtmann et al. 2018) as part of the Detection of RNA Outlier Pipeline DROP v1.0.3 (Yépez et al. 2021). We used the GRCh37 primary assembly, release 29, of the GENCODE project as the reference genome (Frankish et al. 2019). RNA expression outliers were defined as those with FDR of 0.1 or lower.

Results

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). 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. 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

P-values assuming either a normal distribution or a Student's t-distribution fitted for each protein. For each sample, the *P*-values are then corrected for multiple testing. Outliers for protein-sample combinations are defined as those below a predefined FDR threshold (Methods).

Proteomics data often contain missing values which complicates model fitting (Brenes et al. 2019, Phua, Lim and Goh 2022). We filtered out proteins with more than 30% missing values across an entire dataset. The remaining missing values were handled by setting the missing values to the geometric mean of the protein intensities in the input of the autoencoder and ignoring them in the loss computation and *P*-value calculations.

We considered two options to find the optimal encoding dimension of the latent space of the autoencoder. The first one conducts a grid search across several values of the encoding dimension and selects the one that optimizes the recovery of artificially injected outliers. The second option applies to linear autoencoders without covariates 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).

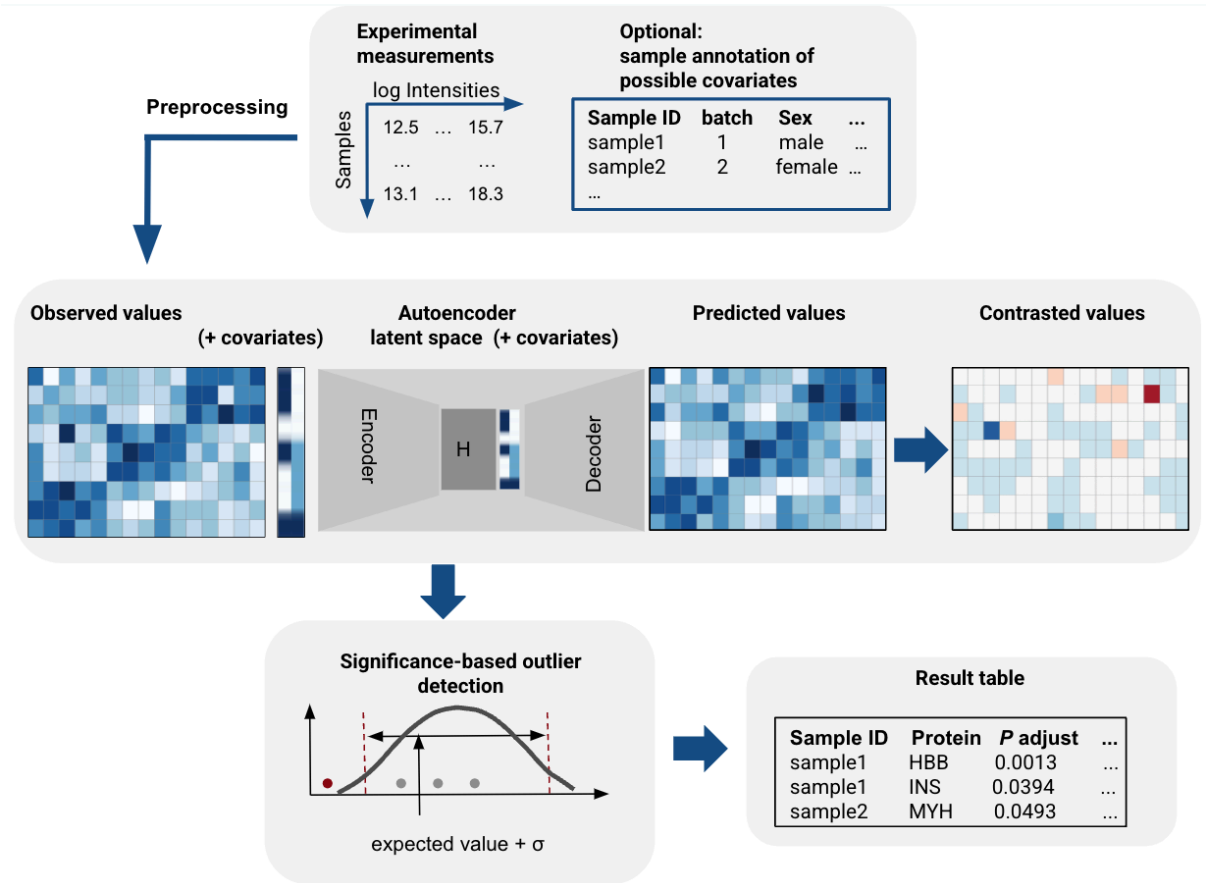

Fig. 1: Schematic overview of the PROTRIDER outlier detection approach. PROTRIDER takes a protein intensity matrix from a quantitative proteomics experiment as input, and optionally covariates such as batch, and fits a conditional autoencoder to account for known and unknown biological and

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.

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, 7,060 proteins were quantified (Supplementary Fig. S1). 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 cancer cell lines. The two panels were analyzed separately because they differed in the number of proteins detected, cancer 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 (Supplementary Fig. S2).

Strong batch effects were observed in the raw log-transformed protein intensities between samples for all three datasets (Fig. 2A, Supplementary Fig. S3A, C). 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 (Supplementary Fig. S3A, C).

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. We refer to this model as ‘PROTRIDER (covariates)’. We compared it to an autoencoder that did not have access to information about potential covariates during model fitting. To this end, we use a linear autoencoder that we initialized using truncated singular value decomposition and with a dimension determined with the OHT approach. The parameters of this model were further tuned (Methods). We named this model ‘PROTRIDER (OHT)’. 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 50, 11, and 5 to be optimal for ‘PROTRIDER (covariates)’ on the mitochondrial disorder dataset, NCI60, and CRC65, respectively. For the ‘PROTRIDER (OHT)’ method, encoding dimensions of 41, 8, and 6 were obtained, respectively. Remarkably, the OHT-proposed dimensions landed in the performance plateau of the optimal dimensions found by grid search (Supplementary Fig. S4).

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 with OHT and to 0.14 ± 0.081 for PROTRIDER with covariates. These results show that despite not having access to covariates, PROTRIDER with OHT succeeded in capturing the

covariation. Similar results were obtained on the tumor cell line panel datasets (Supplementary Fig. S3B, D).

When modeling the residuals with protein-specific normal distributions, we observed an excess of one-sided *P*-values 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 (Methods). The *P*-values 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, C).

We exemplify PROTRIDER outlier calls with the data, 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, appears as a downregulated outlier in the individual OM02364 (FDR = 0.05). This is evident from the deviation of the observed intensity compared to the PROTRIDER expectation by 61-fold (Fig. 2D, E), which is 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'). Moreover, we considered an approach based on the differential expression test limma (Ritchie et al. 2015), which we ran for each sample against the others. This differential expression method allowed for the adjustment of known covariates. At an FDR of 0.1, PROTRIDER typically reported 1 to 2 outliers 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). In contrast, the limma approaches typically yielded larger and more variable numbers of outliers across the datasets (median up to 7 outliers per sample depending on the dataset and the inclusion of covariates, Fig. 2F).

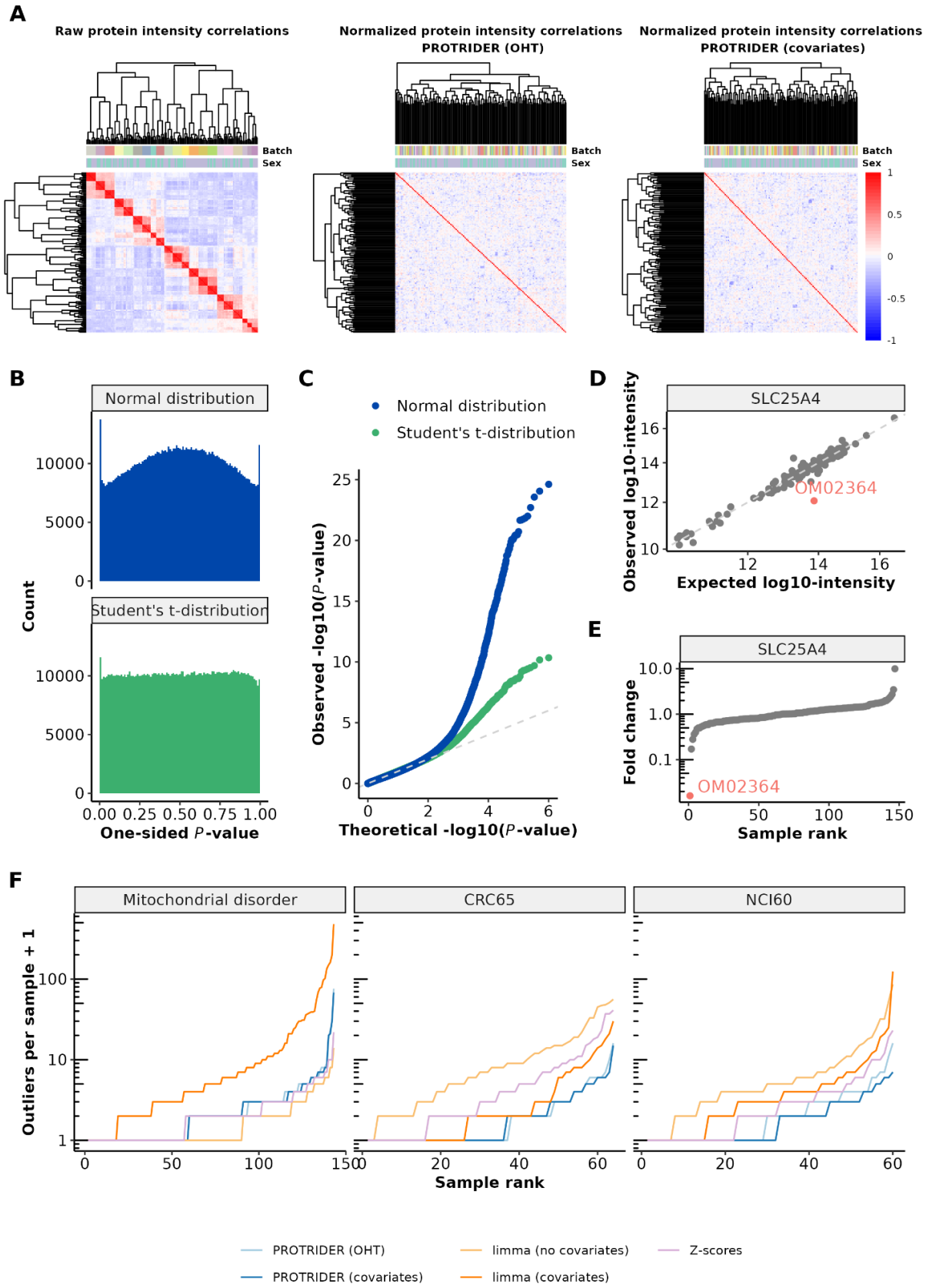

Fig. 2: PROTRIDER accounts for known and hidden covariates and typically detects a limited number of outliers per sample. A) Heatmaps of sample-sample correlations of protein log-transformed intensities on the mitochondrial disorder dataset before (left) and after the

PROTRIDER with OHT (middle) and with autoencoder correction using sex and batch as covariates (right). **B)** Histogram of *P*-values from the normal distribution (blue) and from the Student's t-distribution (green) obtained after the PROTRIDER (OHT) correction on the mitochondrial disorder dataset. **C)** Quantile-Quantile plot comparing observed $-\log_{10}(P)$ values to their theoretical quantiles under the null (uniform distribution). **D)** Observed against expected log-transformed protein intensities for the gene SCL25A4 highlighting the protein underexpression outlier on individual OM02364. **E)** Sorted fold changes of protein intensities (ratio of observed and fitted values) for the gene SCL25A4 highlighting individual OM02364. **F)** Sorted number of protein abundance outliers per sample obtained by the four methods: PROTRIDER (covariates, dark blue), PROTRIDER (OHT, light blue), the limma-based approaches with (dark orange) and without covariates (light orange), and the Z-score-based (violet) approach on the three datasets (facets).

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. 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, Cheng et al. 2023, Rentzsch et al. 2019). 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 limma-based and Z-scores approaches, 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 1,000 top-ranked underexpression outliers detected by PROTRIDER with covariates, 11.6% ([9.7, 13.8] 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. In comparison, only 7.4% ([5.9, 9.2] 95% confidence interval) of the outliers called by the limma-based differential expression approach including covariates harbored at least one such rare variant. The Z-score approach, along with a limma-based differential expression without covariates, showed substantially lower proportions (never exceeding 5%). The superiority of PROTRIDER was observed independently of the rank cutoff and for all three considered variant categories (Fig. 3A). Remarkably, PROTRIDER (OHT) reached nearly equal and not significantly fewer proportions than PROTRIDER with covariates despite not having had access to the covariates. 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 were globally lower compared to the mitochondrial dataset. Perhaps, weaker enrichments were obtained in these datasets because of the smaller sample size (65 and 60 compared to 143), because tumor cell lines are more genetically heterogeneous than patient fibroblasts, or because of a combination of these factors. Nonetheless, considering the significant calls only, PROTRIDER reported much fewer outliers than the limma-based approach (51 vs. 206 on NCI60, 55 vs. 102 on CRC65) and maintained in the case of NCI60 a reasonably high proportion of rare variants likely causing protein abundance aberration compared to the limma-based approach (8.8% vs 7.6% at rank 500, Fig. 3B, C). Altogether, this benchmark using independent evidence from rare genetic variants indicates that PROTRIDER in either setting improves the detection of genuine 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 (Supplementary Fig. S5). Therefore, we used one-layer autoencoders for all the following analyses.

As another independent type of benchmarking data, we considered gene expression outliers called 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 (Kusnadi et al. 2022, Wang et al. 2018). Nevertheless, the proportions of RNA-seq outliers obtained on the same samples can be used to compare different protein abundance outlier callers. 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 three methods which performed similarly (Fig. 3D). For overexpression, no enrichment for RNA-seq outliers was found by the limma approach without covariates nor by the Z-scores approach, while the remaining methods performed similarly and significantly higher than random (Fig. 3D).

Taken together, these results show that PROTRIDER is well suited for the task of aberrant protein abundance detection, outperforming the Z-scores approach and 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. While the two PROTRIDER approaches that we compared performed very similarly in all the benchmarks, the version that uses the grid search approach for finding the optimal encoding dimension and explicitly provided relevant covariates achieved slightly better results when benchmarking for rare-protein disrupting variants. However, 'PROTRIDER (OHT)' is substantially less computationally intensive (~1 minute for the PROTRIDER (OHT) autoencoder fit and ~27 minutes for the approach including covariates and grid search on a Rocky Linux 8.10-based server with 128 physical cores and 512 GB available memory utilizing 16 cores in parallel and 32GB per PROTRIDER run). Hence, 'PROTRIDER (OHT)' is likely a valuable option for many users. When prior knowledge of strong sources of variation in the data is available, we nonetheless recommend investigating the conditional autoencoder PROTRIDER method.

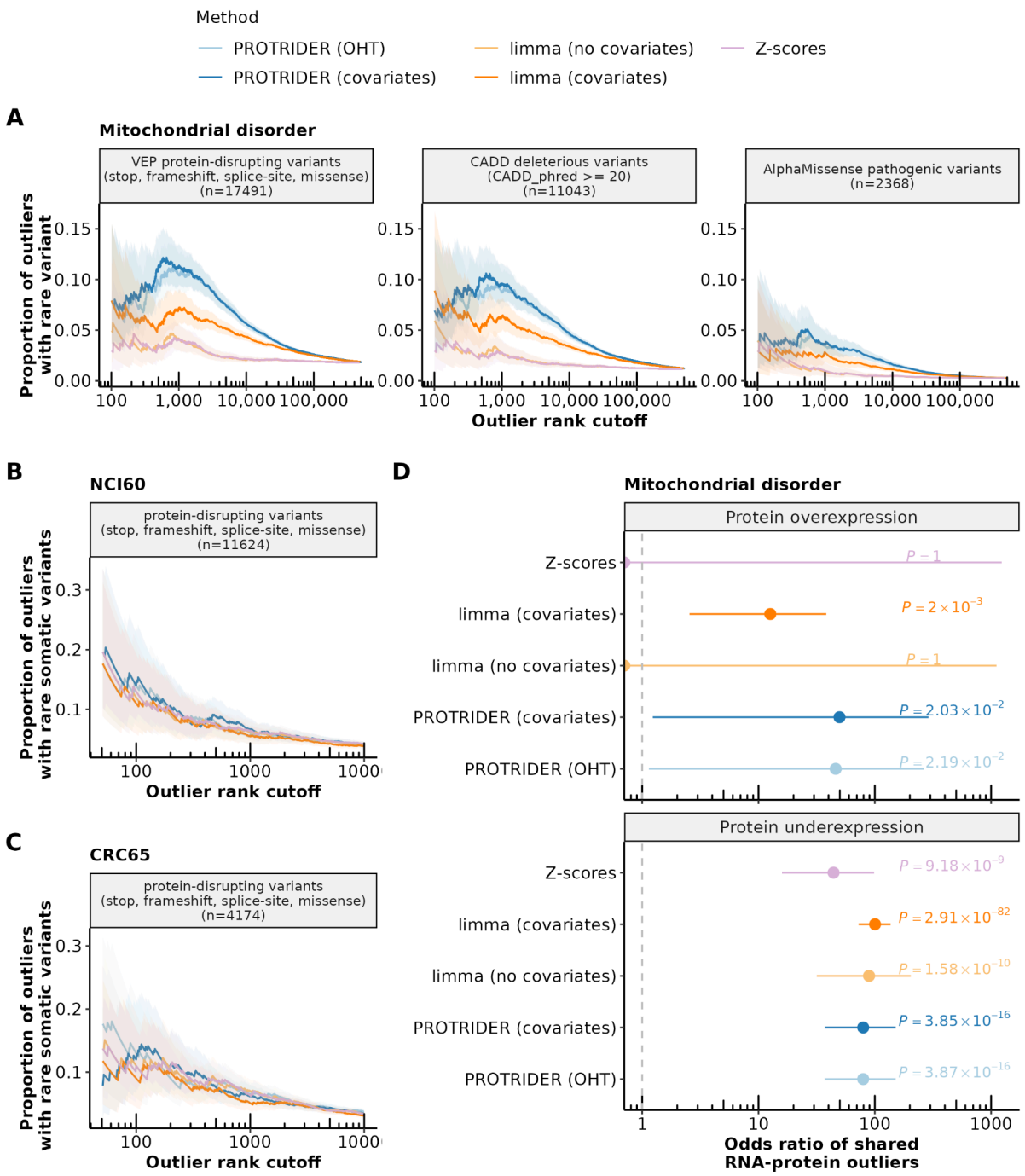

Fig. 3: PROTRIDER outperforms baseline approaches on rare variant benchmarks.
A) Proportion of outliers with at least one rare variant on the mitochondrial disorder dataset for underexpression outliers calls from PROTRIDER (covariates, light blue), PROTRIDER (OHT, blue), the limma-based method with covariates (orange) and without covariates (light orange), and the z-score-based method with PCA correction (dark violet) and without correction (light violet) on 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-C)** Same as A) but for the two cancer 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 in the mitochondrial disorder dataset and their 95% confidence intervals (Fisher's test) for the two PROTRIDER methods, the two limma-based approaches, and the Z-scores approach.

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. Considering PROTRIDER with covariates, we found that 21% of the underexpression protein abundance outliers had at least one such rare variant (Fig. 4A). As expected, stop, frameshift, and splice site variants showed the strongest enrichments (Fig. 4B) and explained overall 10.8% of the underexpression outliers (Fig. 4A). Stronger enrichments were found for PROTRIDER with covariates, but very similar results were obtained when using PROTRIDER OHT (Fig. 4A, B). No significant enrichments for these variants were found, as expected, for overexpression outliers (Supplementary Fig. S6).

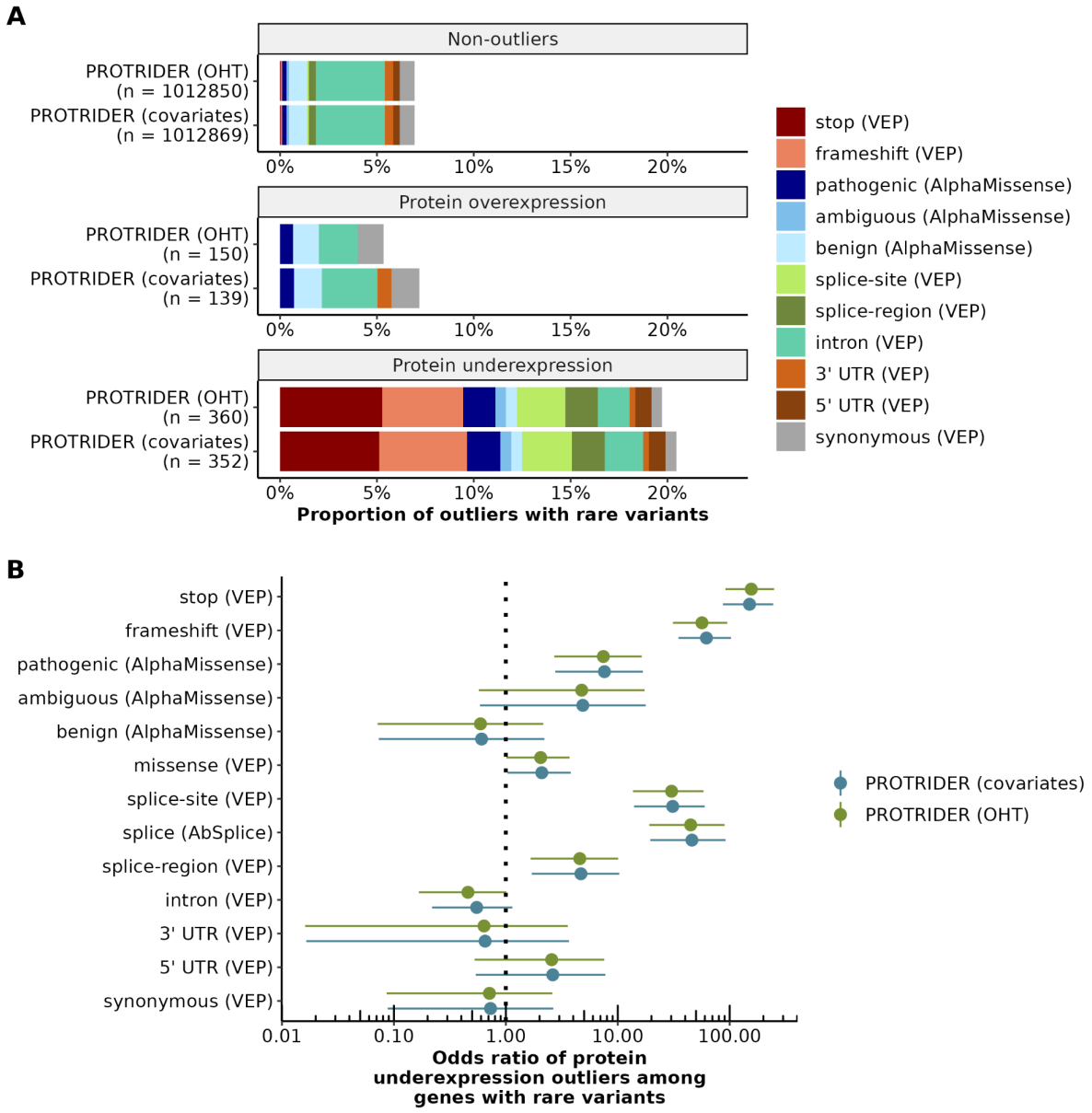

Fig. 4: Genetic determinants of protein expression outliers. A) Proportions of non-outliers, protein overexpression, and protein underexpression outliers with a rare variant detected in the same gene for two PROTRIDER versions: PROTRIDER (OHT) and PROTRIDER with covariates sex and batch in the mitochondrial 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.

We further found a strong enrichment for AlphaMissense pathogenic variants (Fig. 4A, 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 = 7.6) had stronger enrichment than the ambiguous (odds ratio = 4.8) and benign (odds ratio = 0.6) categories (Fig. 4B).

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 underexpression outliers detected with PROTRIDER outperformed baseline methods based on differential expression and simple Z-scores. In the three investigated datasets from rare disease and cancer cell lines, linear models outperformed non-linear, multi-layer autoencoders. Moreover, we found that model residuals were better captured with a Student's t-distribution than with a Gaussian distribution. Finally, we showed that variants predicted by AlphaMissense to be pathogenic, in contrast to the benign predictions, were enriched among underexpression outliers.

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 is able to provide functional evidence for rare variants of uncertain evidence. This is true for stop and frameshift variants but also for missense variants, which represent the most 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.

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 as 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 PROTIDER

overexpression calls were enriched for RNA-seq overexpression calls, indicating that the method does not suffer from major issues on the overexpression side.

Missing values are observed frequently in mass spectrometry-based proteomics datasets, especially those created with a data-dependent acquisition approach (Brenes et al. 2019, Karpievitch, Dabney and Smith 2012, Kong et al. 2022, Jin et al. 2021, Ahlmann-Eltze and Anders 2020). Proper handling of missing values is therefore crucial. However, completely removing missing values from the analysis may not be the preferred solution, as there is also information contained in the missing values since they are generally more likely to occur in low-abundant proteins (Kong et al. 2022). Currently, PROTRIDER does not report *P*-values for missing values, which likely overlooks some outliers that reduce the protein abundance towards zero, and we observe stronger enrichments for proteins without any missing values. This hints towards the need for more sophisticated handling of missing values. One such way could be to move toward variational autoencoders specifically designed for handling incomplete data (Nazábal et al. 2020, Collier, Nazabal and Williams 2021) or approaches that model dropout (Ahlmann-Eltze and Anders 2020).

Another limitation is the number of proteins captured in a typical proteomics experiment. The datasets we considered reliably quantified only approximately 7,000 to 8,000 proteins, which falls within the typical range of such assay (Sun et al. 2023, Frejno et al. 2020, Birhanu 2023). This number is lower than what is usually captured in RNA-seq gene expression analyses, where about 12,000 to 15,000 genes are detected depending largely on the probed tissue (Yépez et al. 2022). As such, the ability of PROTRIDER to detect aberrant protein abundances caused by genomic alterations depends on the protein in question being expressed in the sampled tissue, as well as the ability of the experimental protocol to quantify protein levels reliably across a wide range of proteins. To unlock the full potential of protein outlier detection with PROTRIDER, efforts to expand the number of detected proteins in the experiment will be of great value.

Furthermore, PROTRIDER currently only provides *P*-values at the protein level. Additionally including *P*-values for each protein complex is likely to be helpful, as in some cases whole complexes are affected by an aberrant protein, and many or all of the proteins forming the complex exhibit reduced abundances, which might not be significant on their own (Kremer et al. 2017, Kopajtich et al. 2021). Moreover, providing *P*-values at the peptide level could provide finer resolution in outlier detection by distinguishing between changes driven by specific peptide variants, such as those resulting from post-translational modifications or alternative splicing events, which could have functional implications.

While we developed PROTRIDER with proteomics measurements in mind, other continuous quantities such as abundance measures from metabolomics or lipidomics (Doğan 2024, Alseekh et al. 2021, Salihovic et al. 2023, Züllig and Köfeler 2021), can also be quantified using mass spectrometry-based experiments and may exhibit similar statistical properties. Assuming autoencoder correction and statistical calibration of the model also hold for those data, this would provide an even broader field of application for PROTRIDER to other mass spectrometry-based measurements.

Data availability

No original data was created for this study. The RNA-seq and proteomics data of the mitochondrial dataset is available at GHGA (Dataset ID: GHGAMCS41579681109068). The proteomics data of the NCI60 and CRC65 cell line panels were obtained through PRIDE accessions PXD013615 and PXD005354 as noted in (Frejno et al. 2020).

Competing interests

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.

Acknowledgments

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.

Funding

This study 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.); 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 via the EVUK program ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria [to I.S. and J.G.].

Author contribution

Conceptualization: JG. Data Curation Management: IS, DKA, VAY. Formal Analysis: IS, DKA, VAY. Resources: DS, HP. Software: IS, DKA, SL. Supervision: VAY, JG. Validation: IS, DKA, VAY. Visualization: IS, DKA, VAY, JG. Writing—original draft preparation: IS, DKA, VAY, JG. Writing—review & editing: all authors.

References

Abaan, Ogan D., Polley, Eric C., Davis, Sean R., Zhu, Yuelin J., Bilke, Sven, Walker, Robert L., et al., 'The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems Pharmacology', *Cancer Research*, 73/14 (2013), 4372–82

Ahlmann-Eltze, Constantin, and Anders, Simon, 'proDA: Probabilistic Dropout Analysis for Identifying Differentially Abundant Proteins in Label-Free Mass Spectrometry', *bioRxiv*, 2020, 661496

Alseekh, Saleh, Aharoni, Asaph, Brotman, Yariv, Contrepois, Kévin, D'Auria, John, Ewald, Jan, et al., 'Mass Spectrometry-Based Metabolomics: A Guide for Annotation, Quantification and Best Reporting Practices', *Nature Methods*, 18/7 (2021), 747–56

Benjamini, Yoav, and Yekutieli, Daniel, 'The Control of the False Discovery Rate in Multiple Testing under Dependency', *The Annals of Statistics*, 29/4 (2001), 1165–88

Birhanu, Alemayehu Godana, 'Mass Spectrometry-Based Proteomics as an Emerging Tool in Clinical Laboratories', *Clinical Proteomics*, 20/1 (2023), 32

Brechtmann, Felix, Mertes, Christian, Matusevičiūtė, Agnė, Yépez, Vicente A., Avsec, Žiga, Herzog, Maximilian, et al., 'OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data', *The American Journal of Human Genetics*, 103/6 (2018), 907–17

Brenes, Alejandro, Hukelmann, Jens, Bensaddek, Dalila, and Lamond, Angus I., 'Multibatch TMT Reveals False Positives, Batch Effects and Missing Values *', *Molecular & Cellular Proteomics*, 18/10 (2019), 1967–80

Çelik, Muhammed Hasan, Gagneur, Julien, Lim, Ryan G., Wu, Jie, Thompson, Leslie M., and Xie, Xiaohui, 'Identifying Dysregulated Regions in Amyotrophic Lateral Sclerosis through Chromatin Accessibility Outliers', *Human Genetics and Genomics Advances*, 5/3 (2024) <https://www.cell.com/hgg-advances/abstract/S2666-2477(24)00057-5> [accessed 20 August 2024]

Cheng, Jun, Novati, Guido, Pan, Joshua, Bycroft, Clare, Žemgulytė, Akvilė, Applebaum, Taylor, et al., 'Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense', *Science*, 381/6664 (2023), eadg7492

Collier, Mark, Nazabal, Alfredo, and Williams, Christopher K. I., 'VAEs in the Presence of Missing Data' (2021) <http://arxiv.org/abs/2006.05301> [accessed 20 August 2024]

Cummings, Beryl B., Marshall, Jamie L., Tukiainen, Taru, Lek, Monkol, Donkervoort, Sandra, Foley, A. Reghan, et al., 'Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing', *Science Translational Medicine*, 9/386 (2017), eaal5209

'DepMap 24Q2 Public' (2024) <https://plus.figshare.com/articles/dataset/DepMap_24Q2_Public/25880521/1> [accessed 30 August 2024]

Doğan, Halef Okan, 'Metabolomics: A Review of Liquid Chromatography Mass Spectrometry-Based Methods and Clinical Applications', *Turkish Journal of Biochemistry*, 49/1 (2024), 1–14

Frankish, Adam, Diekhans, Mark, Ferreira, Anne-Maud, Johnson, Rory, Jungreis, Irwin, Loveland, Jane, et al., 'GENCODE Reference Annotation for the Human and Mouse Genomes', *Nucleic Acids Research*, 47/D1 (2019), D766–73

Frejno, Martin, Meng, Chen, Ruprecht, Benjamin, Oellerich, Thomas, Scheich, Sebastian, Kleigrewe, Karin, et al., 'Proteome Activity Landscapes of Tumor Cell Lines Determine Drug Responses', *Nature Communications*, 11/1 (2020), 3639

Frésard, Laure, Smail, Craig, Ferraro, Nicole M., Teran, Nicole A., Li, Xin, Smith, Kevin S., et al., 'Identification of Rare-Disease Genes Using Blood Transcriptome Sequencing and Large Control Cohorts', *Nature Medicine*, 25/6 (2019), 911–19

Gavish, Matan, and Donoho, David L., ‘The Optimal Hard Threshold for Singular Values Is 4/\sqrt 3’, *IEEE Transactions on Information Theory*, 60/8 (2014), 5040–53

Hock, Daniella H., Caruana, Nikeisha J., Semcesen, Liana N., Lake, Nicole J., Formosa, Luke E., Amarasekera, Sumudu S. C., et al., ‘Untargeted Proteomics Enables Ultra-Rapid Variant Prioritization in Mitochondrial and Other Rare Diseases’, *medRxiv*, 2024, 2024.08.06.24311318

Jenkinson, Garrett, Li, Yang I, Basu, Shubham, Cousin, Margot A, Oliver, Gavin R, and Klee, Eric W, ‘LeafCutterMD: An Algorithm for Outlier Splicing Detection in Rare Diseases’, *Bioinformatics*, 36/17 (2020), 4609–15

Jin, Liang, Bi, Yingtao, Hu, Chenqi, Qu, Jun, Shen, Shichen, Wang, Xue, et al., ‘A Comparative Study of Evaluating Missing Value Imputation Methods in Label-Free Proteomics’, *Scientific Reports*, 11/1 (2021), 1760

Jumper, John, Evans, Richard, Pritzel, Alexander, Green, Tim, Figurnov, Michael, Ronneberger, Olaf, et al., ‘Highly Accurate Protein Structure Prediction with AlphaFold’, *Nature*, 596/7873 (2021), 583–89

Karczewski, Konrad J., Francioli, Laurent C., Tiao, Grace, Cummings, Beryl B., Alföldi, Jessica, Wang, Qingbo, et al., ‘The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans’, *Nature*, 581/7809 (2020), 434–43

Karpievitch, Yuliya V., Dabney, Alan R., and Smith, Richard D., ‘Normalization and Missing Value Imputation for Label-Free LC-MS Analysis’, *BMC Bioinformatics*, 13/16 (2012), S5

Kingma, Diederik P., and Ba, Jimmy, ‘Adam: A Method for Stochastic Optimization’ (2017) <http://arxiv.org/abs/1412.6980> [accessed 21 January 2025]

Kong, Weijia, Hui, Harvard Wai Hann, Peng, Hui, and Goh, Wilson Wen Bin, ‘Dealing with Missing Values in Proteomics Data’, *PROTEOMICS*, 22/23–24 (2022), 2200092

Kopajtich, Robert, Smirnov, Dmitrii, Stenton, Sarah L., Loipfinger, Stefan, Meng, Chen, Scheller, Ines F., et al., ‘Integration of Proteomics with Genomics and Transcriptomics Increases the Diagnostic Rate of Mendelian Disorders’, *medRxiv*, 2021 <https://www.medrxiv.org/content/early/2021/07/03/2021.03.09.21253187>

Kremer, Laura S., Bader, Daniel M., Mertes, Christian, Kopajtich, Robert, Pichler, Garwin, Iuso, Arcangela, et al., ‘Genetic Diagnosis of Mendelian Disorders via RNA Sequencing’, *Nature Communications*, 8/1 (2017), 15824

Kusnadi, Eric P., Timpone, Clelia, Topisirovic, Ivan, Larsson, Ola, and Furic, Luc, ‘Regulation of Gene Expression via Translational Buffering’, *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1869/1 (2022), 119140

Labory, Justine, Le Bideau, Gwendal, Pratella, David, Yao, Jean-Elisée, Ait-El-Mkadem Saadi, Samira, Bannwarth, Sylvie, et al., ‘ABEILLE: A Novel Method for ABerrant Expression Identification empLoying Machine LEarning from RNA-Sequencing Data’, *Bioinformatics*, 38/20 (2022), 4754–61

Love, Michael I., Huber, Wolfgang, and Anders, Simon, ‘Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2’, *Genome Biology*, 15/12 (2014), 550

McLaren, William, Gil, Laurent, Hunt, Sarah E., Riat, Harpreet Singh, Ritchie, Graham R. S., Thormann, Anja, et al., 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17/1 (2016), 122

Mertes, Christian, Scheller, Ines F., Yépez, Vicente A., Çelik, Muhammed H., Liang, Yingjiqiong, Kremer, Laura S., et al., 'Detection of Aberrant Splicing Events in RNA-Seq Data Using FRASER', *Nature Communications*, 12/1 (2021), 529

Nazábal, Alfredo, Olmos, Pablo M., Ghahramani, Zoubin, and Valera, Isabel, 'Handling Incomplete Heterogeneous Data Using VAEs', *Pattern Recognition*, 107 (2020), 107501

Phua, Ser-Xian, Lim, Kai-Peng, and Goh, Wilson Wen-Bin, 'Perspectives for Better Batch Effect Correction in Mass-Spectrometry-Based Proteomics', *Computational and Structural Biotechnology Journal*, 20 (2022), 4369–75

Rentzsch, Philipp, Witten, Daniela, Cooper, Gregory M, Shendure, Jay, and Kircher, Martin, 'CADD: Predicting the Deleteriousness of Variants throughout the Human Genome', *Nucleic Acids Research*, 47/D1 (2019), D886–94

Ritchie, Matthew E., Phipson, Belinda, Wu, Di, Hu, Yifang, Law, Charity W., Shi, Wei, et al., 'Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies', *Nucleic Acids Research*, 43/7 (2015), e47

Roumeliotis, Theodoros I., Williams, Steven P., Gonçalves, Emanuel, Alsinet, Clara, Del Castillo Velasco-Herrera, Martin, Aben, Nanne, et al., 'Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells', *Cell Reports*, 20/9 (2017), 2201–14

Salihovic, Samira, Lamichane, Santosh, Hyötyläinen, Tuulia, and Orešič, Matej, 'Recent Advances towards Mass Spectrometry-Based Clinical Lipidomics', *Current Opinion in Chemical Biology*, 76 (2023), 102370

Salkovic, Edin, Abbas, Mostafa M., Belhaouari, Samir Brahim, Errafii, Khaoula, and Bensmail, Halima, 'OutPyR: Bayesian Inference for RNA-Seq Outlier Detection', *Journal of Computational Science*, 47 (2020), 101245

Salkovic, Edin, Sadeghi, Mohammad Amin, Baggag, Abdelkader, Salem, Ahmed Gamal Rashed, and Bensmail, Halima, 'OutSingle: A Novel Method of Detecting and Injecting Outliers in RNA-Seq Count Data Using the Optimal Hard Threshold for Singular Values', *Bioinformatics*, 39/4 (2023), btad142

Scheller, Ines F., Lutz, Karoline, Mertes, Christian, Yépez, Vicente A., and Gagneur, Julien, 'Improved Detection of Aberrant Splicing with FRASER 2.0 and the Intron Jaccard Index', *The American Journal of Human Genetics*, 110/12 (2023), 2056–67

Segers, Alexandre, Gilis, Jeroen, Van Heetvelde, Mattias, De Baere, Elfride, and Clement, Lieven, 'Juggling Offsets Unlocks RNA-Seq Tools for Fast Scalable Differential Usage, Aberrant Splicing and Expression Analyses', *bioRxiv*, 2023, 2023.06.29.547014

Smail, Craig, and Montgomery, Stephen B., 'RNA Sequencing in Disease Diagnosis', *Annual Review of Genomics and Human Genetics*, 25/Volume 25, 2024 (2024), 353–67

Sun, Benjamin B., Chiou, Joshua, Traylor, Matthew, Benner, Christian, Hsu, Yi-Hsiang,

Richardson, Tom G., et al., ‘Plasma Proteomic Associations with Genetics and Health in the UK Biobank’, *Nature*, 622/7982 (2023), 329–38

Tyanova, Stefka, Temu, Tikira, and Cox, Juergen, ‘The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics’, *Nature Protocols*, 11/12 (2016), 2301–19

Vialle, Ricardo A., de Paiva Lopes, Katia, Bennett, David A., Crary, John F., and Raj, Towfique, ‘Integrating Whole-Genome Sequencing with Multi-Omic Data Reveals the Impact of Structural Variants on Gene Regulation in the Human Brain’, *Nature Neuroscience*, 25/4 (2022), 504–14

Wagner, Nils, Çelik, Muhammed H., Hölzlwimmer, Florian R., Mertes, Christian, Prokisch, Holger, Yépez, Vicente A., et al., ‘Aberrant Splicing Prediction across Human Tissues’, *Nature Genetics*, 55/5 (2023), 861–70

Wang, Kai, Li, Mingyao, and Hakonarson, Hakon, ‘ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data’, *Nucleic Acids Research*, 38/16 (2010), e164

Wang, Sidney H., Hsiao, Chiaowen Joyce, Khan, Zia, and Pritchard, Jonathan K., ‘Post-Translational Buffering Leads to Convergent Protein Expression Levels between Primates’, *Genome Biology*, 19/1 (2018), 83

Yépez, Vicente A., Gusic, Mirjana, Kopajtich, Robert, Mertes, Christian, Smith, Nicholas H., Alston, Charlotte L., et al., ‘Clinical Implementation of RNA Sequencing for Mendelian Disease Diagnostics’, *Genome Medicine*, 14/1 (2022), 38

Yépez, Vicente A., Mertes, Christian, Müller, Michaela F., Klaproth-Andrade, Daniela, Wachutka, Leonhard, Frésard, Laure, et al., ‘Detection of Aberrant Gene Expression Events in RNA Sequencing Data’, *Nature Protocols*, 16/2 (2021), 1276–96

Zecha, Jana, Satpathy, Shankha, Kanashova, Tamara, Avanessian, Shayan C., Kane, M. Harry, Clauser, Karl R., et al., ‘TMT Labeling for the Masses: A Robust and Cost-Efficient, In-Solution Labeling Approach *[S]’, *Molecular & Cellular Proteomics*, 18/7 (2019), 1468–78

Züllig, Thomas, and Köfeler, Harald C., ‘High Resolution Mass Spectrometry in Lipidomics’, *Mass Spectrometry Reviews*, 40/3 (2021), 162–76

Supplemental Figures

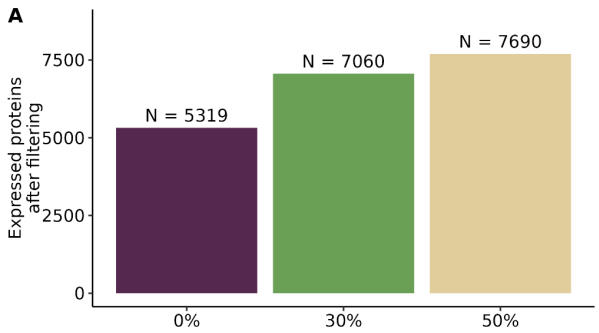

Fig. S1: Expressed proteins and rare variants per category on the mitochondrial disorder dataset. A) Number of expressed proteins after filtering out proteins with too many missing values per protein, for three cutoffs on the maximal allowed percentage of missing values: no missing values (0%, purple), at most 30% missing values (green), and at most 50% missing values per protein (beige).

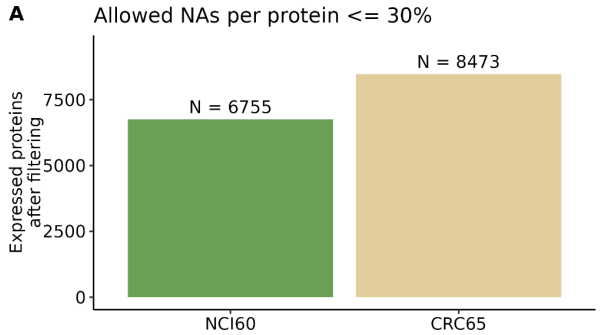

Fig. S2: Expressed proteins and rare variants per category on the cancer cell line datasets.
A) Number of expressed proteins after filtering out proteins with more than 30% missing values per protein, for the cell line panels NCI60 (green) and CRC65 (beige).

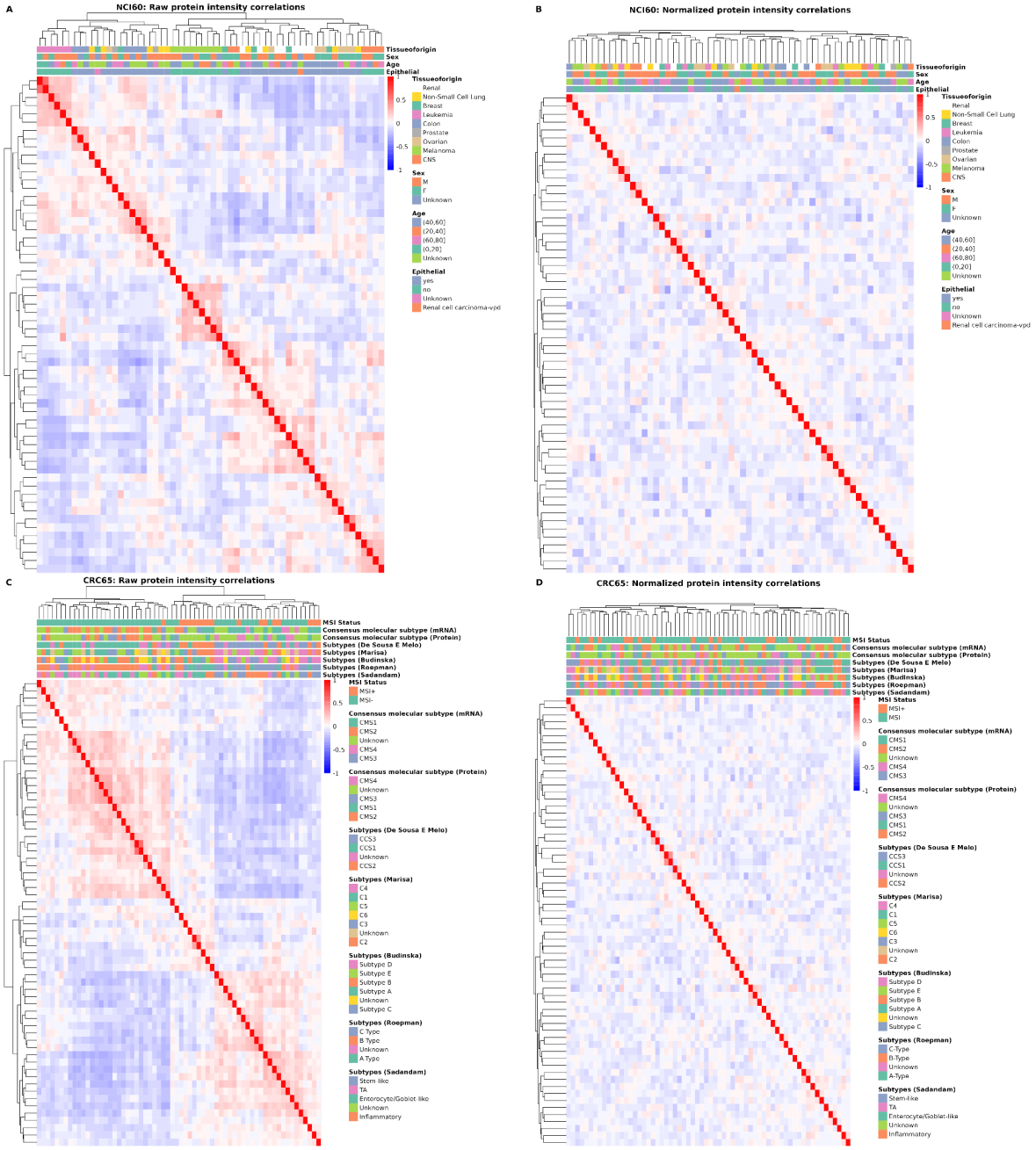

Fig. S3: Sample-sample correlations before and after autoencoder on the cancer cell line panels. A-B) Heatmaps of sample-sample correlations of protein log-transformed intensities before (A) and after the PROTRIDER autoencoder correction for the PROTRIDER version using OHT for finding the optimal encoding dimension (B) on the NCI60 cell line panel data. **C-D)** Same as (A) and (B), respectively, but for samples from the cell line panel CRC65. The within-batch pairwise sample correlation was reduced from 0.15 ± 0.11 (mean ± standard deviation) to 0.079 ± 0.062 for NCI60 and from 0.11 ± 0.074 to 0.05 ± 0.04 for CRC65'

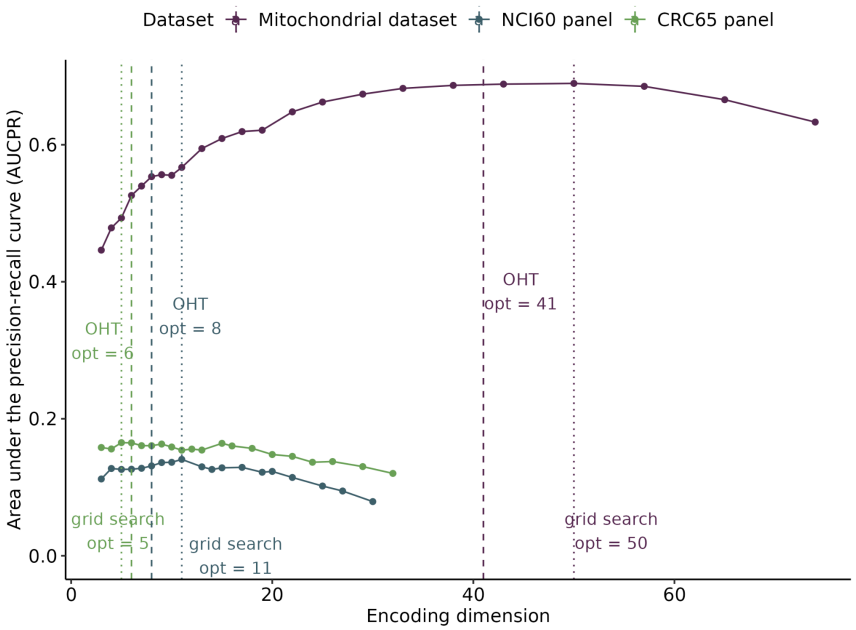

Fig. S4: Optimal encoding dimensions found by PROTRIDER with and without provided covariates. The obtained area under the precision-recall curve of recovering injected outliers is plotted against the tested encoding dimensions. The dotted lines indicate the optimal encoding dimension found for each dataset with the grid search approach and the OHT approach.

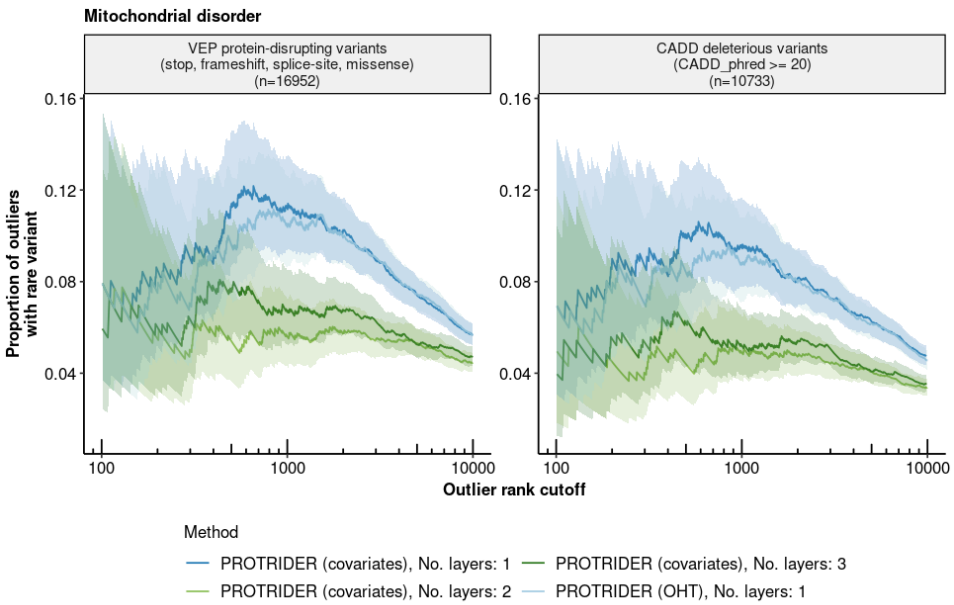

Fig. S5: Linear autoencoder model outperforms multi-layer autoencoder models. Proportion of outliers with at least one rare variant on the mitochondrial disorder dataset for underexpression outliers calls from PROTRIDER (covariates) with different number of layers and, PROTRIDER (OHT) with two sets of rare variant categories as ground truth proxies: i) VEP stop, frameshift, direct split-site, and missense variants and ii) CADD deleterious variants (PHRED score ≥ 20).

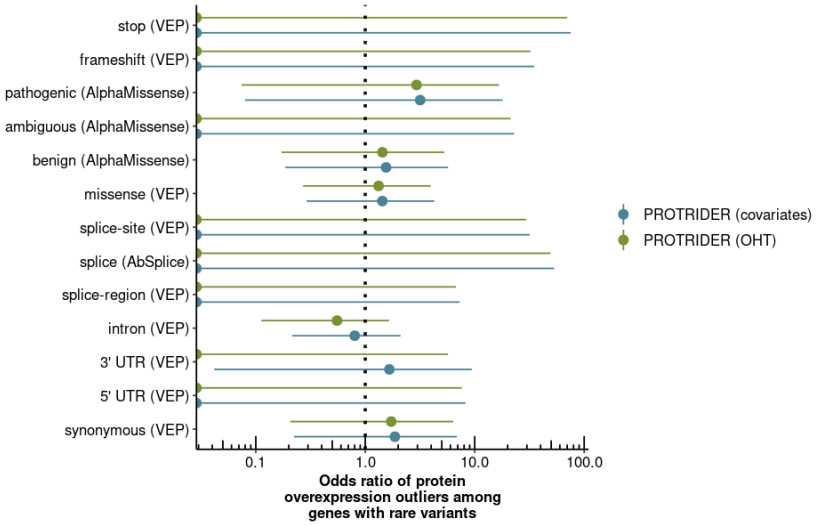

Fig. S6: Genetic determinants of protein overexpression expression outliers. Odds ratios and their 95% confidence intervals (Fisher's test) of the enrichment of the proportion for each variant category among overexpression outliers compared to the background proportion of the non-outliers.