Received: 21 April 2023 | Revised: 13 September 2023 | Accepted: 5 October 2023

DOI: 10.1002/jex2.120

RESEARCH ARTICLE

Proteome encoded determinants of protein sorting into extracellular vesicles

Katharina Waury[1] | **Dea Gogishvili**[1] | **Rienk Nieuwland**[2,3] | **Madhurima Chatterjee**[4] | **Charlotte E. Teunissen**[5] | **Sanne Abeln**[1,6]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2]Laboratory of Experimental Clinical Chemistry, Department of Clinical Chemistry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[3]Vesicle Observation Centre, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[4]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[5]Neurochemistry Laboratory, Department of Clinical Chemistry, Amsterdam Neuroscience, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[6]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Correspondence
Sanne Abeln, Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, 1081HV, The Netherlands.
Email: s.abeln@vu.nl

Funding information
H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 860197; EU Joint Programme – Neurodegenerative Disease Research, Grant/Award Number: bPRIDE; Health Holland, Grant/Award Number: LSHM20106; ZonMw, Grant/Award Number: 73305095007; Alzheimer's Drug Discovery Foundation; Selfridges Group Foundation; Alzheimer Nederland; Alzheimer's Association

Abstract

Extracellular vesicles (EVs) are membranous structures released by cells into the extracellular space and are thought to be involved in cell-to-cell communication. While EVs and their cargo are promising biomarker candidates, sorting mechanisms of proteins to EVs remain unclear. In this study, we ask if it is possible to determine EV association based on the protein sequence. Additionally, we ask what the most important determinants are for EV association. We answer these questions with explainable AI models, using human proteome data from EV databases to train and validate the model. It is essential to correct the datasets for contaminants introduced by coarse EV isolation workflows and for experimental bias caused by mass spectrometry. In this study, we show that it is indeed possible to predict EV association from the protein sequence: a simple sequence-based model for predicting EV proteins achieved an area under the curve of 0.77 ± 0.01, which increased further to 0.84 ± 0.00 when incorporating curated post-translational modification (PTM) annotations. Feature analysis shows that EV-associated proteins are stable, polar, and structured with low isoelectric point compared to non-EV proteins. PTM annotations emerged as the most important features for correct classification; specifically, palmitoylation is one of the most prevalent EV sorting mechanisms for unique proteins. Palmitoylation and nitrosylation sites are especially prevalent in EV proteins that are determined by very strict isolation protocols, indicating they could potentially serve as quality control criteria for future studies. This computational study offers an effective sequence-based predictor of EV associated proteins with extensive characterisation of the human EV proteome that can explain for individual proteins which factors contribute to their EV association.

KEYWORDS
biomarkers, extracellular vesicles, human proteome, machine learning, post-translational modification

1 | INTRODUCTION

Extracellular vesicles (EVs) are a heterogeneous group of lipid-delimited vesicles that are released by cells into the extracellular space (Borges et al., 2013; Yáñez-Mó et al., 2015). EVs have been observed across a manifold of cell types and all domains

Katharina Waury and Dea Gogishvili contributed equally to this work.

of life highlighting their universal biological importance (van Niel et al., 2018). On the basis of their function, cargo, size and excretion pathways, EVs may be divided into three main types: exosomes, microvesicles and apoptotic bodies (Yáñez-Mó et al., 2015). Biomolecules associated with EVs are strong biomarker candidates as they can be isolated from easily accessible body fluids and provide insight into the state of the donor cells (Gámez-Valero et al., 2019). The clinical potential of EV biomarkers has been elucidated for several pathologies (Camino et al., 2021; Miranda et al., 2010; Whiteside, 2015), especially for neurodegenerative diseases (Gámez-Valero et al., 2019; Watson et al., 2019). Thus, there is a strong need to better understand the underlying mechanisms of protein association with EVs. However, the investigation of EVs is hampered by the difficulties of correctly isolating and concentrating EVs out of complex body fluids. Although experimental guidelines are being established (Gandham et al., 2020; Lötvall et al., 2014; Théry et al., 2018), there is a continued discussion about the optimal workflow, and many published studies do not meet the minimal criteria of EV enrichment and isolation (Théry et al., 2018). Thus, the standardisation of isolation and identification techniques is still a major hurdle, despite its importance for successful EV biomarker development (Coumans et al., 2017; Witwer et al., 2013). EV studies are prone to include non-EV contaminants during sample collection, isolation, concentration and characterisation of EVs (Coumans et al., 2017). For instance, isolation kits, such as ExoQuick, might enrich EVs, but result in various contaminants, including antibodies and polymers (Théry et al., 2018). It has been suggested that more than 70% of all particles isolated from blood plasma are not EVs and the major contaminants are lipoprotein particles (Sodar et al., 2016; Welton et al., 2015) along with platelets (Karimi et al., 2018). For studies investigating the presence and absence of unique proteins in EVs, it is of particular importance that the number of unique contaminants in the data is as low as possible.

Continuous advancements in research have substantially contributed to EV proteome characterisation. Vesiclepedia (Pathan et al., 2018) and ExoCarta (Keerthikumar et al., 2016; Simpson et al., 2012) are two major manually curated databases compiling identified EV cargo. While these two databases provide a rich information source, they must be treated with caution and analysed carefully considering major types of artefacts likely to be present in the data. It is crucial to be aware of how reported proteins may be associated with EVs. Proteins listed in these databases can be (i) inside the vesicles ('cargo'), (ii) on the outside ('protein corona'), as well as (iii) bound to the membrane. Thus, we use the term EV association and collectively refer to such proteins as EV proteins. For biomarker research, all types of EV association may be relevant as the presence of a biomarker candidate within the EV lumen or membrane might explain its limited or missing detection, for example, using immunoassays (Waury et al., 2022). Knowledge of EV association can provide directions for experimental analysis, that is, the need for EV isolation and disruption techniques prior to protein detection. Additionally, the systematic analysis of EV proteins might provide valuable insights into their characteristics and thereby can increase our understanding of the cell's sorting mechanisms and the function of EVs.

While there have been numerous attempts to predict subcellular (Kumar & Dhanda, 2020), as well as extracellular matrix (Liu et al., 2020; Zhao et al., 2019) localisation of proteins based on machine learning approaches, prediction of EV localisation has been pursued much less. Ras-Carmona et al. attempted the prediction of protein secretion by EVs but limited their study to exosome cargo proteins. They reported an area under the curve (AUC) of 0.76 ± 0.03 using dipeptide composition features (Ras-Carmona et al., 2021). In this study, we focus on EV association predictions for the human proteome. In addition, we put an emphasis on the explainability of the machine learning models, both to reveal general sorting trends, as well as to identify potential mechanisms for individual proteins.

There are two primary aims of this study: (1) To ascertain the possibility of predicting EV association based on amino acid sequence using machine learning; (2) to investigate if the presence or absence of a specific human protein in EVs is associated with the sequence, physicochemical and structural features of this protein, as well as post-translational modification (PTM) annotations. We analysed publicly available human EV protein data and corrected these for a potential detection bias by excluding proteins not recognized in mass spectrometry (MS) studies and proteins identified by unreliable EV isolation workflows. We constructed a wide variety of informative protein properties as input for the machine learning models; all these properties could either be calculated directly from the protein sequence or were derived from curated database annotations. These handcrafted features allowed us to interpret the machine learning model predictions, and link them to potential EV sorting mechanisms. This study offers an effective sequence-based predictor of 'a ticket to a bubble ride' (Anand et al., 2019) and an extensive and systematic characterisation of the human EV proteome.

2 | RESULTS

In this study, we aimed to answer if the prediction of EV association is a feasible task. To obtain annotations suitable to train such a predictor, we designed a data curation workflow combining several resources and filtering steps as seen in Figure 1 and described in detail in Section 4. Subsequently, we trained a machine learning model on a comprehensive set of features (Tables S1 and S2), and identified the properties most characteristic of EV and non-EV proteins both at a global level and for individual proteins. The list of curated EV proteins is provided in the supplement (Table S3).

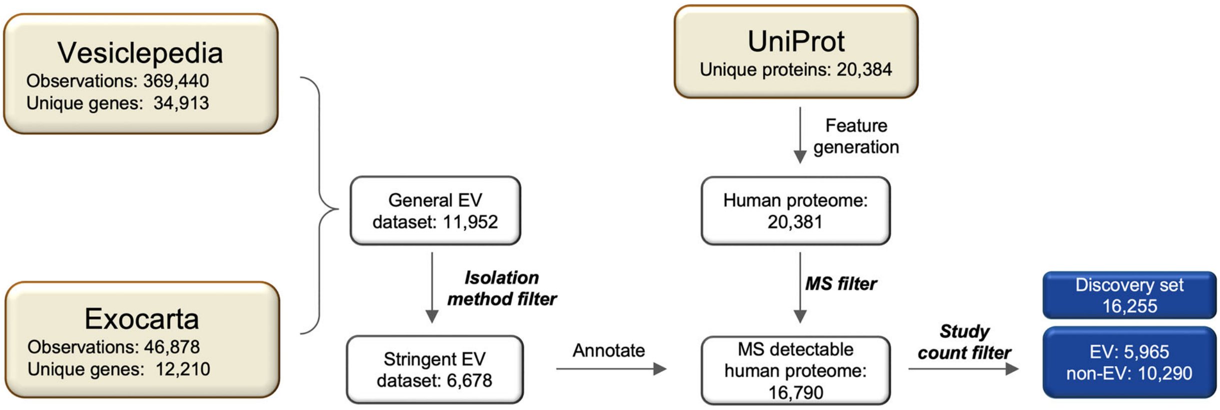

FIGURE 1 Data curation workflow. Squares represent datasets with remaining entries. Three datasets from databases Vesiclepedia, ExoCarta and UniProt are coloured in khaki. Unique proteins from Vesiclepedia and ExoCarta were merged to construct a General EV dataset, and proteins identified by unreliable isolation workflows were removed to obtain the Stringent EV dataset. Sequence-based features as well as annotations were generated for each protein in the human proteome. Human proteins not detectable by MS were removed by the MS filter. All unique MS-detectable human proteins were annotated regarding their EV association using the stringent EV dataset. Lastly, rarely detected EV proteins (count ≤ 2) were removed from the dataset entirely resulting in the EV-annotated discovery set (blue). EV, extracellular vesicle; MS, mass spectrometry.

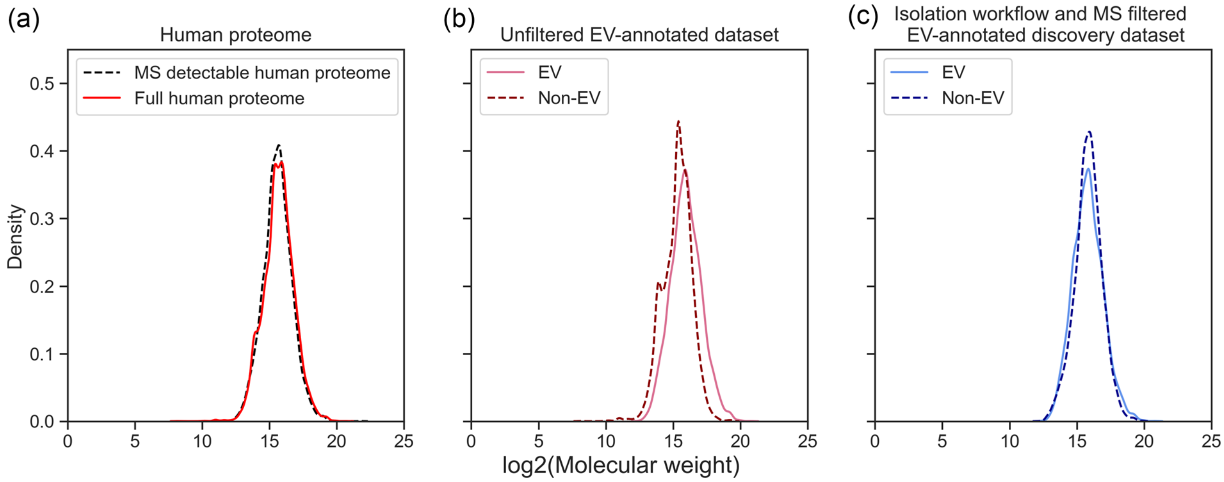

FIGURE 2 Density plots of log2-transformed molecular weight across the human proteome and EV-annotated datasets. (a) Distribution of log2-transformed molecular weight of the MS-detectable human proteome compared to the full human proteome. MS struggles to detect low molecular weight proteins of the human proteome. (b) The molecular weight densities of EV and non-EV proteins in the unfiltered EV annotated dataset highlight the discrepancy in molecular weight between EV and non-EV proteomes. (c) The much more similar molecular weight distribution of EV and non-EV group show how the MS filter step diminishes the experimental bias introduced by MS. EV, extracellular vesicle; MS, mass spectrometry.

2.1 | EV proteomics data requires extensive data filtering steps

Several filtering steps were incorporated into the EV data curation workflow to reduce the likelihood of contaminants and any systematic bias introduced by mass spectrometry (MS); this workflow led to the EV-annotated discovery dataset (Figure 1). Additionally, we examined EV-annotated datasets that skip some of the filtering steps to determine what properties are truly connected to EV sorting processes and which might be a manifestation of biases; the data curation for these datasets is shown in Figure S1.

While MS is most suitable for the broad detection of proteins, it creates experimental data biased towards proteins of higher molecular weight and lower isoelectric point (Klont et al., 2018). As most EV database entries have been identified by MS, we wished to investigate and correct for bias introduced by MS. We analysed the effect of the MS filter, as described in Section 4, by comparing molecular weight distributions in the datasets with and without this filtering step. MS clearly struggles to detect low molecular weight proteins within the human proteome as especially low molecular weight proteins are missing in the MS-detectable proteome compared to the full human proteome (Figure 2a). This confirms previous results (Klont et al., 2018). This bias is strongly pronounced in the unfiltered EV-annotated dataset (Figure 2b). Here, the non-MS-detectable subset of the human proteome is annotated as non-EV which explains the low molecular weight peaks in the non-EV proteome. When applying the

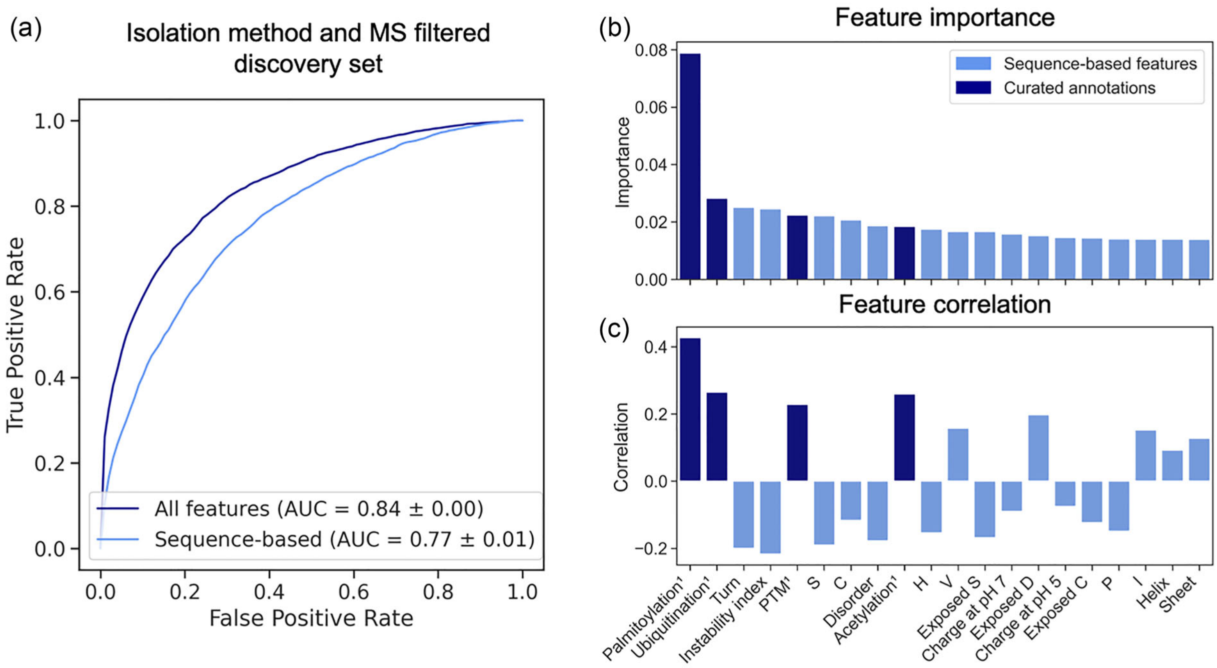

FIGURE 3 Performance of RF model and feature importance analysis. (a) ROC curves and AUC display the performance of the RF classifiers using sequence-based features (light blue) and sequence-based features and curated annotations (dark blue). (b) Bar plots show the Gini importance of the top 20 features for EV prediction as well as the correlation of these features with the EV class (c). Note that (b) and (c) share the same labels. PTMs, stability, structure, and polarity differentiate EV and non-EV proteins. Dark blue features are curated annotations. AUC, area under the curve; C, cysteine; D, aspartic acid; EV, extracellular vesicle; H, histidine; I, isoleucine; P, proline; PTM, post-translational modification; RF, random forest; ROC, receiver operating characteristic; S, serine; V, valine.

MS filter, and thus, using only the MS-detectable subset of the human proteome, the discrepancy in molecular weight between EV and non-EV associated proteins is much smaller (Figure 2c), illustrating that the constructed filter successfully corrects for technical bias introduced by MS. While we only show the effects of the filtering step in terms of protein size, we assume this filter will also help to correct any additional unknown MS detection biases.

2.2 | Prediction of EV associated proteins is feasible

To investigate if there is a signal in a protein's sequence that determines its EV association, we constructed machine learning models to predict this property. The sequence properties of a protein are the input features for the model; the output is a prediction if the protein is EV associated or not. Two different random forest (RF) models were trained on the discovery dataset: (i) incorporating only sequence-based features; (ii) incorporating sequence-based features and curated annotations. Figure 3a displays their respective receiver operating characteristic (ROC) curves and AUC scores.

The model trained on the discovery set achieved a ROC-AUC of 0.77 ± 0.01, which increased further to 0.84 ± 0.00 when incorporating curated annotations as input. The superior model performance, when also including curated annotations, suggests that this type of information is valuable for the correct prediction of EV association and cannot be compensated for by sequence-based predictions alone.

We also trained models on the alternative datasets that excluded some of the filter steps (Figure S2). The slightly better performance of the classifier which is not trained on an MS-filtered dataset indicates that this classifier effectively (also) predicts MS detectability.

2.3 | Various features are important for prediction of EV association

To gain insight into the differences between EV and non-EV-associated human proteins, we examined the protein features most decisive for the correct classification of the proteins. Figure 3b displays the Gini importance of the 20 most important features for EV association of the model trained on the discovery set. Figure 3c indicates these 20 features if the correlation with the EV class is positive or negative. To compare which trends are identical and diverging in datasets with fewer filter steps, a heat map ranking all features across the three trained models was generated (Figure S3).

PTM annotations comprise evidently many of the most important features. Notably, all PTM types investigated here are positively correlated with the EV class indicating enrichment of PTMs in EV-associated proteins. Especially palmitoylation shows high correlation (Figure 3c) and is by far the most important feature used by the model to predict EV association (Figure 3b). Sequence-based protein features considered important for prediction include predicted structure (Turn, Disorder, Helix, Sheet), stability (Instability index) and charge (Charge at pH 5 and 7, Exposed D) (Figure 3b).

The abundances of less common lipidation PTMs, that is, prenylation, myristoylation and GPI-anchors, are shown Figure S4a. We were especially interested in those PTMs as palmitoylation, also a lipidation modification, was found to be so strongly enriched in the human EV proteome. Other rare PTMs have been linked to EV sorting (Anand et al., 2019; Campanella et al., 2015; Moreno-Gonzalo et al., 2014): citrullination, ISGylation, nitration and NEDDylation (Figure S4b). While for these PTM types the number of associated proteins is extremely limited, we could still show enrichment in the EV group. Less frequent PTMs were not important for EV classification by the RF because of their scarce absolute annotation numbers (Figure S3); nevertheless, these are biologically interesting.

The feature importance of curated PTM annotations was much higher than the importance of predicted PTMs by MusiteDeep. Hence, the differences between EV and non-EV proteins in terms of PTM annotations cannot be fully captured by state-of-the-art prediction methods that are purely sequence-based (Figure S3).

Additionally, we compared the feature importance results with a ‘random’ model, which was trained on shuffled and thus meaningless EV labels, and as expected, the importance of the protein properties evened out and lost relevance. A comprehensive overview of every feature is provided in the supplement containing each feature’s adjusted *p*-value regarding the comparison of EV and non-EV protein groups and their correlation with the EV class. For additional statistical comparison, we provide the feature importance of the original model and a model trained on shuffled (i.e., meaningless) EV labels (Table S4).

2.4 | Observations hold true in a high confidence EV set

In order to validate the determinants for EV association found by the machine learning models (Figure 3b) and simple correlation analysis (Figure 3c), we devised sets of EV annotations with different degrees of reliability.

We combined three recent studies from three different biological fluids which used effective isolation techniques, reducing incorrect EV annotations, for example, due to lipoproteins or platelets (Karimi et al., 2018), resulting in a set of high confidence EV annotations. In addition, we devised a low confidence EV annotation set using older studies, for which we would expect contaminants to be included caused by insufficient isolation workflows.

Figure 4 shows the comparison between physicochemical and structural characteristics (Figure 4a,b), as well as PTM annotations (Figure 4c,d), of the high confidence validation EV proteins (HC EV), the proteins annotated by low confidence EV studies (LC EV), and the EV and non-EV datasets of the discovery set which was used for training and testing our prediction model. The associations of the investigated features that we identified in our EV dataset become even stronger in the high confidence EV dataset, while identified EV-specific signals become diluted in the low confidence EV set. All selected sequence-based features consistently show a clear trend: the difference in each feature between non-EV the EV sets increases, the higher the confidence of the data (Figure 4a,b), suggesting that we identified true EV-specific determinants in our EV discovery dataset. The confirmation of the PTM properties of EV proteins is evident as well, especially for palmitoylation and nitrosylation (Figure 4c). In fact, in the high confidence EV set, over half of the proteins contain a palmitoylation site, confirming the prevalence of this EV sorting mechanism. On the other hand, the PTM signals become diluted in the low confidence dataset, probably due to many falsely reported discoveries in the experimental data. These results suggest that EV association determinants that could be confirmed in the high confidence dataset, for example, palmitoylation, may be used for quality control purposes of experimental data.

Predicting on a set of novel EV proteins from the EVpedia database, our model correctly predicts 41.27% of these proteins as EV associated. Thus, the model sensitivity is lower on this independent validation set compared to the discovery set. Note, however, that false positive proteins could be present in this novel set of EV proteins as no filtering based on error-prone isolation workflows was possible as was done for the discovery dataset. Importantly, when we only included EVpedia proteins with a higher number of occurrences, the sensitivity of the model improved, illustrating that our model performs better for higher confidence EV proteins of EVpedia (Figure S5).

2.5 | Functional characterisation of frequently detected EV proteins

In addition to the general characteristics of the EV proteome, we were interested in proteins that are often identified in EV studies and their properties. To functionally characterize EV-associated proteins, we selected 478 unique human proteins from Vesiclepedia with occurrences (counts) in at least 30 different studies. Pathway enrichment analysis revealed proteins most frequently detected in EVs to be strongly associated with the ribosome. Notably, ribosomal proteins have been previously reported to be

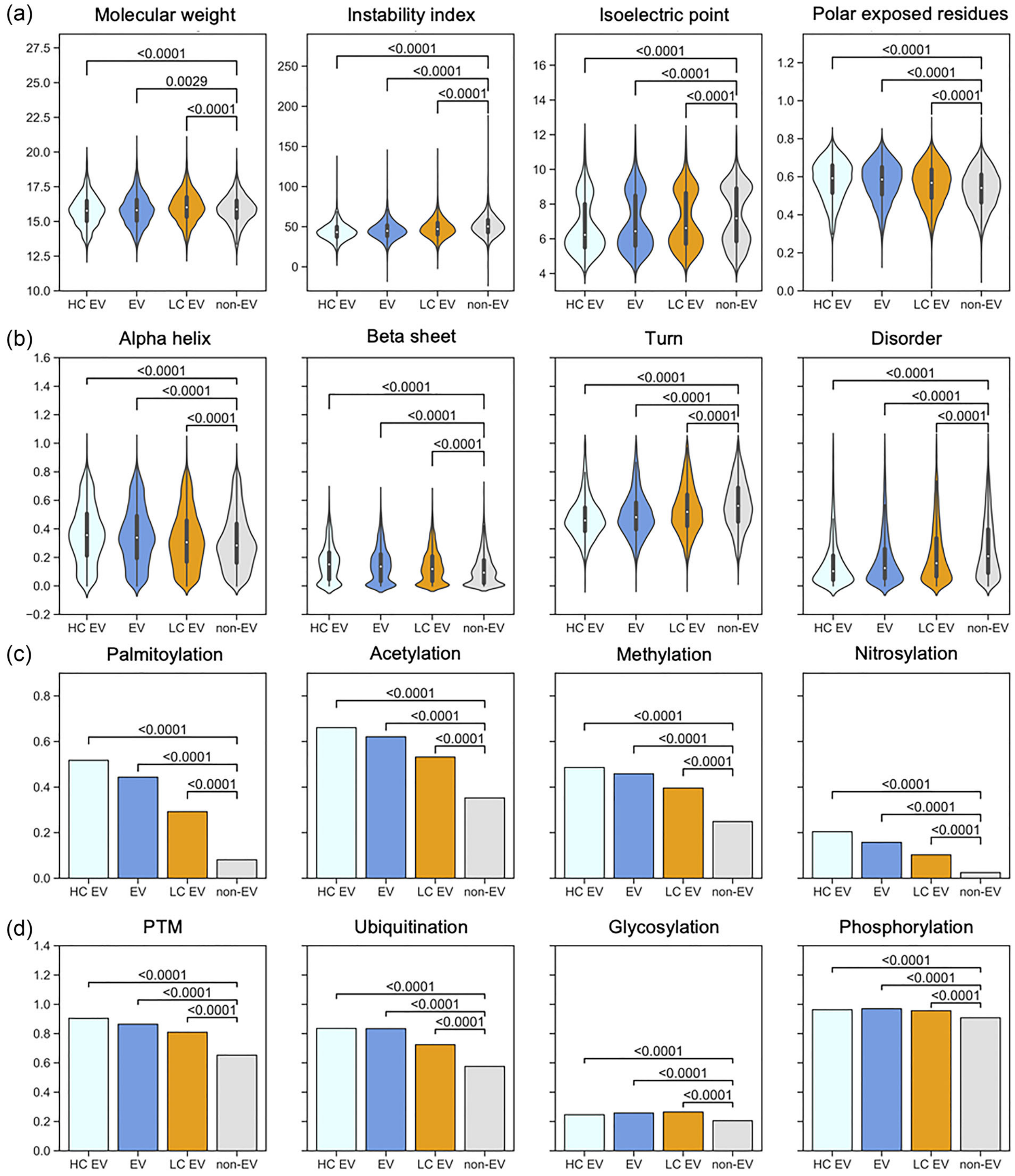

FIGURE 4 Features in the high and low confidence EV sets. Proteins in the high confidence EV dataset, which was constructed from three recent studies show a similar distribution of physicochemical and structural properties as the EV protein set of the discovery dataset. For many features, the discrepancy with the non-EV group becomes more distinct. Furthermore, the low confidence dataset (orange) which contains EV proteins identified in older studies dilutes the observed signal compared to the EV protein set probably due to many falsely included contaminants. *p*-values are displayed in the plots. EV, extracellular vesicle; HC, high confidence; LC, low confidence.

enriched in exosomes (Dozio & Sanchez, 2017). Other enriched pathways include the phagosome, cell-cell communication and immune response pathways, including viral carcinogenesis, *Escherichia coli* infection, and antigen presentation (Figure S6).

2.6 | SHAP plots illustrate the model's decision-making for case examples

After establishing the most important characteristics of EV proteins on a global level, we aimed to illustrate why the model predicts a protein to be EV associated (or not) for individual proteins. Note that this type of information cannot be derived

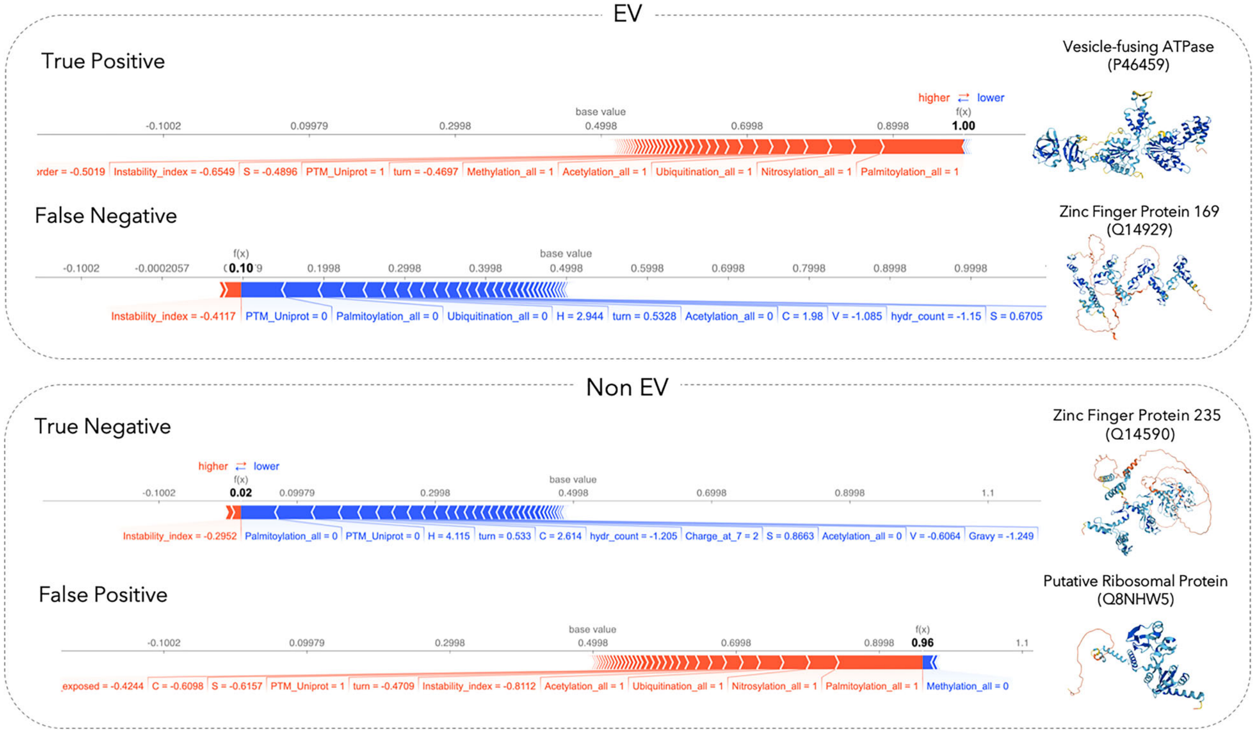

FIGURE 5 Shapley value analysis of case examples. SHAP plots for local interpretability are shown for correctly predicted proteins (i.e., true positive, true negative), as well as proteins for which the model prediction and the annotation from our data curation workflow do not agree with each other (i.e., false positive, false negative) from our test set. We chose examples in which the predictor is very certain if the individual protein is EV associated or not. Each SHAP plot displays a set of SHAP values that explain for each individual protein which features contributed to the model's prediction. Features in red contribute to the prediction being higher (i.e., EV associated), and features in blue decrease the predicted score (i.e., non-EV). Protein structures shown here are predicted by AlphaFold (Jumper et al., 2021).

directly for experimental annotations of EV association, and hence the prediction model can provide additional insight into possible sorting mechanisms associated with a specific protein.

Using SHAP values, the features that cause the model to predict EV association for a specific protein were analysed (Figure 5). SHAP plots show the importance of PTM presence for each case and their contribution to the predicted label.

Of interest is the case study of a false positively predicted ribosomal protein (UniProt ID: Q8NHW5). Based on the properties this protein follows the general trend of the EV proteome: it is heavily post-translationally modified and is stable, leading the model to predict the protein annotated as non-EV to be confidently classified as an EV-associated protein. Investigation of this protein in our data revealed that while this protein is reported in multiple EV studies according to Vesiclepedia (Pathan et al., 2018), its Gene (Entrez) ID could not be mapped to its UniProt entry using the UniProt ID mapping tool. Thus, it was removed during the filtering steps despite being a true EV protein. Importantly, this circumstance affected approximately 12% of all proteins reported within the EV databases. Nevertheless, our model could correctly classify it.

3 | DISCUSSION

In this work, we asked if it is possible to predict EV association based on the sequence (and associated annotations) of a protein, and which protein properties contributed to EV association predictions, both on a proteome- and an individual protein level.

An extensive multi-step data curation and filtering workflow led to a comprehensive human EV proteome. It is important, however, to consider that some tissues and fluids might be underrepresented within EV research, and thus the human EV proteome might still not be fully characterised. Further, legitimate EV proteins were lost during ID mapping. While a more elaborate, multi-step mapping workflow might have increased the number of retained EV database entries, we have decided to prioritise an easily reproducible and robust workflow only utilising the UniProt ID mapping tool. Nevertheless, the EV proteome dataset allowed us to discriminate true determinants of EV association from an experimental bias inherent to EV databases. Importantly, we analysed unique proteins regarding their presence or absence in EVs, for which defining clear positives and negatives is essential. We showed that it is indeed possible to predict EV association from amino acid sequences with an AUC of 0.77 in our filtered discovery set (Figure 3a). The performance increased further to 0.84 when incorporating curated annotations.

Feature analysis of the EV and non-EV proteome elucidated that EV proteins are more stable, polar, and structured than non-EV proteins and contain various PTM sites (Figure 3b-c). These observed biological trends held true in a high confidence EV set of recent, high-quality EV studies (Figure 4), confirming the determinants of EV association. Furthermore, we show that the low confidence dataset containing proteins identified in older EV studies mostly weakens the signal (Figure 4) highlighting the importance of data curation steps suggested in this work.

We showed that experimental MS data of EV studies are biased towards larger proteins, and that correction during the data curation process is required. The presented approach to correct protein annotations for the introduced bias of MS is broadly applicable to other proteomics-based studies and databases. As many, especially older studies did not contemplate any standardisation or possible contaminants, spurious annotations are highly likely to be present in EV databases. This bias had to be counteracted by limiting the studies to be included and excluding low-count proteins. Note that because this study only considered the presence or absence of a protein in EVs, the impact of a few contaminating proteins that were not successfully excluded from the EV set is limited.

While the accuracy of the simple RF classifier is encouraging, improved performance is conceivable, for example, by the incorporation of residue-specific instead of solely global features (Klausen et al., 2019; Shah & Ou, 2021). As of yet, there has been no attempt to apply deep learning approaches to this task which is a promising approach to pursue but might be limited by the amount of available training data. Notably, this study suggests that smaller, but stricter datasets lead to more trustworthy classifiers. For future advancements of machine learning applications within the EV field, it would be valuable to explore protein embeddings as these would allow us to gain deeper insights into the sequence motifs that play a pivotal role in EV sorting. This becomes particularly valuable when dealing with biomarkers that exist as truncated versions or cleaved fragments of existing proteins. In our current study, we solely focused on analysing full-length sequences and incorporated multiple known motifs as features. However, none of these motifs emerged as a robust predictor for EV association. Further, a predictor of protein association with specific EV subtypes, that is, exosomes, microvesicles and apoptotic bodies, may be of value considering their differing functions and excretion pathways. However, this endeavour would require a better understanding of subtype-specific markers to provide high-quality datasets to train on.

Several of our findings about EV properties correspond to already established knowledge about these vesicles. While in both EV and non-EV proteins the typical bimodal distribution of the isoelectric point was present (Kiraga et al., 2007), an enrichment of proteins with a lower isoelectric point, that is, of more acidic proteins, in the EV proteome was found in our analysis (Figures 3c and 4). Interestingly, previous work on exosomes has shown that acidic conditions favour the existence and release of those vesicles (Ban et al., 2015; Parolini et al., 2009). Another possible explanation for this preference towards acidity is the relationship between isoelectric point distribution and subcellular localisation. The proteomes of the cytoplasm, lysosome, vacuoles and cytoskeleton have been shown to have an acidic protein composition, while proteins of the plasma membrane and mitochondria are more basic (Kiraga et al., 2007; Kurotani et al., 2019). Our findings could indicate an enrichment of the former subcellular proteomes, and a depletion of the latter within EVs. An overlap of EV and lysosome proteomes is particularly expected as the sorting of proteins into both exosomes and lysosomes is dependent on multivesicular bodies (MVBs) (Kalra et al., 2016). An enrichment of cytoplasmic proteins is also coherent as the biogenesis of both exosomes and microvesicles involves the envelopment and budding of cytosolic components at the MVB or plasma membrane, respectively (Mathieu et al., 2019).

The importance of PTMs in the biogenesis, cargo-loading, and release of EVs has been studied intensively (Ageta & Tsuchida, 2019; Anand et al., 2019; Carnino et al., 2020). Our work supports previous findings about the existence or even abundance of phosphorylation (Gonzales et al., 2008), glycosylation (Gerlach & Griffin, 2016; da Silva et al., 2021) and ubiquitination (Ageta & Tsuchida, 2019; Buschow et al., 2005; Smith et al., 2015) in EV proteins. We also detected a higher fraction of proteins with methylation and acetylation sites despite no previous strong evidence of their relevance for EV sorting. However, these modifications might not be specific to the EV protein sorting process as their correlation became weaker with an increasing number of filtering steps (Figure S3).

Nitrosylation has been linked to synaptic vesicles (Wang et al., 2015; Wang et al., 2016) but its general function remains elusive as it has been implicated in cell survival and death, regulation of protein activity and localization amongst others (Koriyama and Furukawa, 2018; Stomberski et al., 2019). As a clear enrichment of nitrosylation sites was found in the EV proteome (Figure 4), their potential role in protein localization into EVs should be explored. Palmitoylation emerged to be the strongest feature for EV association—with over half of the high confidence EV associated proteins containing a palmitoylation site (Figure 4). Interestingly, palmitoylation has been implicated in recent years in the EV sorting of specific proteins (Flemming et al., 2020; Romancino et al., 2018). Moreover, Mariscal et al. showed in a comprehensive study of the EV palmitoyl-proteome a high abundance of palmitoylated proteins in cancer-derived EVs (Mariscal et al., 2020). The enrichment in the human EV proteome for both palmitoylation, as well as other types of lipidation, may be explained by their involvement in binding proteins to the plasma membrane (Picciotto et al., 2020). Generally, the impact of predicted PTMs is significantly less important for EV predictions. Predicted palmitoylation and ubiquitination appeared much lower in the feature ranking than their annotation analogues (Figure S3). Advancements in correct PTM prediction methods are thus essential to utilise predicted PTMs for the characterisation of the EV proteome.

The functional enrichment analysis results for the most frequently detected EV proteins are in line with what is currently considered as major functions of EVs, namely cell-to-cell communication (Kalra et al., 2016; Paolicelli et al., 2019), signalling via membrane receptors or cell fusion by releasing cargo (van Niel et al., 2018). Moreover, evidence for an immunoregulatory function of EVs is increasing (Robbins & Morelli, 2014) as exosomes were shown to stimulate immune responses by presenting antigens on their surface (Bobrie et al., 2011) or as carriers of tumour and pathogenic antigens (Greening et al., 2015; Walker et al., 2009; Wolfers et al., 2001). Furthermore, a recent investigation of EVs shed light on their involvement in various disease pathologies including metabolic disorders (Huang-Doran et al., 2017) and viral infection (Anderson et al., 2016).

Taken together, our study provides a systematic characterisation of the EV proteome and a machine learning tool for predicting EV association. Moreover, we illustrate how to provide a readout of features for specific proteins. Based on our machine learning approach, we highlight possible protein sorting mechanisms and quality control criteria that can guide future EV experiments and downstream analysis.

4 | METHODS

A data curation workflow incorporating several resources and filtering steps was created to obtain protein sets annotated regarding their presence or absence in EVs (Figure 1). We accessed online databases to identify human proteins associated with EVs. Simultaneously, we generated a dataset of all unique human proteins with various sequence-based features, as well as curated annotations. Several different filtering strategies were applied to receive higher confidence EV and non-EV protein sets. The resulting EV and non-EV annotations were used to train and test machine learning classifiers and analyse the features necessary for the correct classification of EV association.

4.1 | Dataset generation

4.1.1 | EV dataset

The EV proteome was downloaded from the EV databases Vesiclepedia (http://microvesicles.org, Version 4.1) (Pathan et al., 2018) and ExoCarta (http://www.exocarta.org, Version 5) (Keerthikumar et al., 2016; Simpson et al., 2012). Entries were filtered for human protein content, and the ‘counts’, that is, the number of experiments that identified each protein, were calculated. After merging the entries from both databases, 13,648 unique proteins were identified as EV associated (Figure 1). Their unique UniProt ID was determined through the ID mapping tool provided by UniProt. If more than one Uniprot ID was mapped to an Entrez ID, only the first entry was retained. This led to 11,952 (87.57%) mapped EV proteins which comprise the General EV dataset.

4.1.2 | Human proteome dataset

Swiss-Prot reviewed human proteins were extracted from UniProt (Consortium, 2019) (release: 2020 05) containing 20,385 unique proteins. For 20,381 of these proteins, the full feature set could be generated which comprises the Human proteome dataset.

4.1.3 | Filtering steps for the EV-annotated discovery dataset

To generate a model able to predict which proteins are associated with EVs, it is essential to have a training dataset that is as accurately labelled as possible. The steps below describe, how we filtered and annotated our datasets to minimise the experimental and publication bias, resulting in our discovery dataset.

Firstly, we generated an isolation method filter (Figure 1). The filter excludes proteins that were identified through workflows prone to a high false discovery rate. The information provided on Vesiclepedia regarding applied isolation methods was used to include solely proteins identified in experiments that: (1) made no use of ExoQuick or similar ‘high recovery but low purity’ isolation kits; (2) were comprised of at least three different isolation steps that enrich for different properties (e.g., size and density); (3) did not apply ultracentrifugation as the first step in their EV isolation workflow. These requirements were selected based on previous findings regarding the limited adequacy of EV isolation purity of ultracentrifugation (Linares et al., 2015; Stam et al., 2021) and ExoQuick (Tian et al., 2019; Veerman et al., 2021) as a single isolation step, respectively. Additionally, several studies support the superiority of multiple isolation steps (Karimi et al., 2018; Stam et al., 2021; Veerman et al., 2021;

Zhang et al., 2020). A list of still included isolation workflows can be found in the supplement. This approach resulted in 6818 retained EV proteins of which 6678 were mapped to respective UniProt IDs.

Secondly, we created a MS filter (see Figure 1). As MS was predominantly utilized for protein detection in the majority of EV studies, experimental detection bias known to be introduced by MS (Klont et al., 2018) is likely to exist in this data. This is specifically problematic for the proteins we label as non-EV associated. Note that these proteins may simply not have been detected in EV studies because of the limitations of MS as a detection method. Thus, we wished to exclude any proteins that have never been identified in MS experiments from the discovery dataset to counteract potential bias. For this purpose, we accessed ProteomicsDB, an expansive collection of MS studies of the human proteome among others (Lautenbacher et al., 2021; Samaras et al., 2019). We collected all human proteins in this database that have been verifiably detected in previous MS studies. Only proteins that are present in both the ProteomicsDB and the human proteome datasets were retained for further analysis (16,790 proteins). By excluding any entries from the human proteome dataset not found in proteomics studies an MS-detectable human proteome dataset was generated (Figure 1).

Finally, we removed low-count proteins, that is, those found in less than three unique experiments, from the entire dataset as those proteins were considered ambiguous. Thus, these are neither included as EV nor non-EV proteins.

The final discovery dataset contains 5965 and 10,290 EV- and non-EV-annotated proteins, respectively. To examine the effect of the filtering steps described above, we also trained models based on datasets that skip some of the filtering steps, the workflow for generating these alternative discovery sets is provided in Figure S1. Importantly, our labels referring to EV associations are binary (EV/non-EV). Thus, various proteins that are considered to be enriched in EVs, usually the tetraspanins CD9, CD63 and CD81 (Théry et al., 2018) have the same label as other less common EV proteins—EV associated.

4.2 | Feature generation

We generated a range of features to identify properties that significantly differentiate between proteins identified in EVs and those never detected in EVs. In total, 95 features were included for each protein present in our datasets. In order to obtain explainable machine learning models, it is essential to use input features that are easy to interpret, and that can help to reveal potential EV sorting mechanisms, as well as potential biases in the datasets. For this reason, a wide variety of interpretable protein properties was derived directly from the protein sequences; in addition, curated database annotations were added.

4.2.1 | Sequence-based features

Table S1 lists all these features and details their generation. Basic features, such as sequence length, molecular weight, number of residues per amino acid, and polar and hydrophobic amino acid count were calculated directly from the sequence. All amino acid counts were normalised by the length of the protein to obtain amino acid proportions. Sequence length and molecular weight were log2-transformed.

NetSurfP-2.0 (Klausen et al., 2019) was used to predict solvent accessibility, secondary structure and structural disorder for each residue in the protein sequences. The per-residue predictions were then utilised to calculate global features for each protein according to van Gils et al. (van Gils et al., 2022). Residues with relative surface accessibility (RSA) >0.4 were considered exposed. The number of exposed residues per protein, for each amino acid, polar amino acids and hydrophobic amino acids, were calculated and divided by the sum of all exposed amino acids of the respective protein. We calculated the proportion of structural elements (α helices, β sheets, turns, disordered regions), total accessible surface area (TASA), and total hydrophobic surface area (THSA). TASA and THSA were log2-transformed. Relative hydrophobic surface area (RHSA) was derived as the fraction of THSA relative to TASA.

The Bio.SeqUtils.ProtParam Biopython module (Cock et al., 2009) was used to calculate aromaticity (Lobry & Gautier, 1994), instability index (Guruprasad et al., 1990), Gravy (Kyte & Doolittle, 1982), isoelectric point, charge at pH-7 and pH-5. The SoDoPe tool (Bhandari et al., 2020) was used to derive the probability of solubility. A global aggregation propensity score was calculated for each protein. The aggregation propensity score of each amino acid was derived from the experimental work by De Groot et al. (de Groot et al., 2005).

For various domains, specifically coiled-coil, WW domains, RAS profiles, EGF and RRM, the associated PROSITE (De Castro et al., 2006; Hulo et al., 2007; Sigrist et al., 2012) identifiers were extracted (see Table S1). ScanProsite (De Castro et al., 2006) was applied to cross-reference protein sequences and domain motifs against each other and the domain presences were implemented as binary features.

To enable the use of PTMs as a sequence-based feature, MusiteDeep PTM prediction results were incorporated as features (Wang et al., 2018, 2020). A threshold of 0.75 was chosen to include a predicted PTM site and only the highest-scoring PTM was considered at every site. The predicted PTM sites were implemented as a global feature for each protein, indicating if a PTM is present in or absent from the protein sequence.

TMHMM (Krogh et al., 2001) was used to predict transmembrane helices that were included as a binary feature indicating presence or absence.

4.2.2 | Curated annotations

The databases used to create curated annotation and relevant comments are listed in Table S2. Databases of PTM annotations were accessed to annotate the human proteome. Relevant datasets were downloaded from PhosphoSitePlus (Hornbeck et al., 2014), iPTMnet (Huang et al., 2017), Swiss-Palm (Blanc et al., 2019; Pedregosa et al., 2011) and UniProtKB (Consortium, 2019). Annotations for two uncommon PTMs (ISGylation, NEDDylation) were extracted from UniProt via text mining of the comments section of sequence position independent annotations. While some of these databases also provided position-specific modification information, all feature annotations were generated at a protein level indicating if at least one amino acid in the protein sequence is known to be modified regarding each PTM type. UniProt annotations for transmembrane and heat-shock proteins were administered as binary features.

4.3 | Prediction of EV association

To ascertain the possibility to predict the EV association of a protein from the sequence, as well as to determine which features are considered most important for this task, random forest (RF) classification was implemented. The three discovery datasets described in the section Datasets were used and for each of those datasets, two RF models were trained: Using only sequence-based features and using sequence-based features and curated annotations. This resulted in six classifiers.

4.3.1 | Training and interpretation of the random forest

The discovery dataset was split into an 80% training and 20% testing set, resulting in 9544 training and 3251 testing entries. As the classes (EV and non-EV) were unbalanced, undersampling of the class containing a higher number of proteins, that is, the majority class, was utilised to create balanced training datasets. All continuous features were scaled using the robust scaler provided in the scikit-learn Python library (Pedregosa et al., 2011).

The scikit-learn RF model was implemented. To compare the importance of the features for the prediction of EV association the impurity-based feature importance of the RF models, that is, the Gini importance, was assessed. Higher values indicate greater importance for correct classification. Model performance was evaluated using the ROC curve and the AUC score. To illustrate specific cases, Shapley Additive explanations (SHAP) analysis was carried out (Lundberg & Lee, 2017). 'Feature interactions' were calculated to highlight feature combinations used for predictions.

4.3.2 | Validation of EV association determinants

We aimed to validate the protein-encoded determinants that were found to be important features in the machine learning models, by observing trends of these properties in high and low-confidence EV datasets in terms of experimental procedures.

Since present-day EV studies are characterised by superior enrichment and isolation methods, we chose three recent EV studies of different body fluids. The included studies emphasised EV purification with minimal non-EV contamination in their workflows, have been published when minimal EV study requirements were already established by the community, and have not been included in Vesiclepedia or ExoCarta yet. A study using a bottom-up Optiprep density gradient centrifugation protocol identified 1789 proteins in the EV-enriched fraction of urine samples (Dhondt et al., 2020). Karimi et al. focussed on the isolation of plasma EVs without lipoproteins contamination by combining a density cushion with size-exclusion chromatography and were able to identify 1187 proteins (Karimi et al., 2018). Lastly, a study that isolated 1686 proteins from breast cancer cell line MDA-MB-468 cell culture media EVs by precipitation coupled to size exclusion chromatography was included (Martínez-Greene et al., 2021). We combined the identified EV associated proteins into a high-confidence EV dataset containing 3222 proteins, out of which 573 were not annotated as EV proteins in our discovery dataset.

Secondly, we constructed a low-confidence dataset by only selecting proteins from Vesiclepedia identified in EV studies published before the year 2014. This cut-off likely leads to a higher number of spurious protein identification as at this point in time no guidelines for correct EV isolation were widely agreed upon (Lötvall et al., 2014; Théry et al., 2018) while still providing a decent number of studies and thus proteins to be included in the low-confidence set. Out of 9988 unique proteins, 4974 were not annotated as EV proteins in our discovery set.

We hypothesised that the high-confidence dataset will display the same trends of protein properties as the EV proteins of our discovery set; the existence of non- EV contaminants should be limited because of the high isolation standards of the included studies. We expected the signal in the low-confidence dataset to be diluted by contaminants, demonstrating the missing specificity of early EV isolation workflows.

As an additional validation step, we used the trained model to predict on the novel EV proteins of the additional EV database EVpedia (https://evpedia.info/evpedia2_xe/) (Kim et al., 2013). The full set of human EV proteins was downloaded and filtered for proteins that are part of the human proteome. After the removal of the proteins that were already present in our discovery dataset and ambiguous proteins that have been detected in less than three studies from the EVpedia set, 676 novel proteins were left.

4.4 | Statistical analysis

To test for the significance of features, we employed Mann-Whitney U test and Fisher's exact test for continuous and categorical features, respectively. The threshold for significance was set at 0.05 and correction for multiple testing was done using the Benjamini-Hochberg procedure. To test the significance of the feature importance's, we compared our model to a 'random' model. For this, we trained our models with randomly shuffled EV labels and obtained the feature importance of a 'random' model. To provide a comprehensive overview of every feature, we provide a table containing each feature's adjusted *p*-value regarding the comparison of EV and non-EV protein classes, its correlation with the EV class, as well as the feature importance within our trained model, and the random model (trained on shuffled and thus meaningless EV labels) (Table S4).

4.5 | Functional analysis of EV associated proteins

To analyse proteins that are often identified in EVs (top EV proteins), 478 unique human proteins from Vesiclepedia with occurrences (counts) in at least 30 different studies were selected from our discovery set and analysed. GSEApy enrichr tool was used for the pathway enrichment analysis using the KEGG 2019 Human library (Xie et al., 2021).

AUTHOR CONTRIBUTIONS

Katharina Waury: Data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Dea Gogishvili**: Data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Rienk Nieuwland**: Conceptualization; supervision; supporting; validation; writing—review and editing. **Madhurima Chatterjee**: Writing—review and editing. **Charlotte E. Teunissen**: Conceptualization; funding acquisition; project administration; supervision; writing—review and editing. **Sanne Abeln**: Conceptualization; funding acquisition; methodology; project administration; supervision; writing—original draft; writing—review and editing.

ACKNOWLEDGEMENTS

We would like to thank Jose Gavaldá-Garciá for exploring the possibility of this study.

CONFLICT OF INTEREST STATEMENT

Research of Katharina Waury, Dea Gogishvili, Charlotte E. Teunissen, and Sanne Abeln are supported by the European Commission Marie Curie International Training Network, grant agreement No 860197, the MIRIADE project. Charlotte E. Teunissen is supported by JPND (bPRIDE), Health Holland, the Dutch Research Council (ZonMW), Alzheimer Drug Discovery Foundation, The Selfridges Group Foundation, Alzheimer Netherlands, Alzheimer Association. Charlotte E. Teunissen is recipient of ABOARD, which is a public-private partnership receiving funding from ZonMW (#73305095007) and HealthHolland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). More than 30 partners participate in ABOARD. ABOARD also receives funding from Edwin Bouw Fonds and Gieskes-Strijbisfonds. Charlotte E. Teunissen has a collaboration contract with ADx, Neurosciences, Quanterix and Eli Lilly, performed contract research or received grants from AC-Immune, Axon Neurosciences, Biogen, Brainstorm Therapeutics, Celgene, EIP Pharma, Eisai, PeopleBio, Roche, Toyama, Vivoryon. She serves on editorial boards of Medidact Neurologie/Springer, Alzheimer Research and Therapy, Neurology: Neuroimmunology & Neuroinflammation, and is editor of a Neuromethods book Springer.

DATA AVAILABILITY STATEMENT

All code and data related to this work can be found on the relevant GitHub repository for this project: https://github.com/ibivu/ExtracellularVesicles. This includes the complete feature dataset of the discovery set and the lists of EV proteins that were included in each of the different filtering workflows described.

ORCID

Katharina Waury https://orcid.org/0000-0002-8570-7640
Dea Gogishvili https://orcid.org/0000-0001-8809-0861
Rienk Nieuwland https://orcid.org/0000-0002-5671-3400
Madhurima Chatterjee https://orcid.org/0000-0003-1647-3307
Charlotte E. Teunissen https://orcid.org/0000-0002-4061-0837
Sanne Abeln https://orcid.org/0000-0002-2779-7174

REFERENCES

Ageta, H., & Tsuchida, K. (2019). Post-translational modification and protein sorting to small extracellular vesicles including exosomes by ubiquitin and UBLs. *Cellular and Molecular Life Sciences*, *76*, 4829–4848., 12.

Anand, S., Samuel, M., Kumar, S., & Mathivanan, S. (2019). Ticket to a bubble ride: Cargo sorting into exosomes and extracellular vesicles. *Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics*, *1867*(12), 140203.

Anderson, M. R., Kashanchi, F., & Jacobson, S. (2016). Exosomes in viral disease. *Neurotherapeutics*, *13*, 535–546.

Ban, J.-J., Lee, M., Im, W., & Kim, M. (2015). Low pH increases the yield of exosome isolation. *Biochemical and Biophysical Research Communications*, *461*, 76–79.

Bhandari, B. K., Gardner, P. P., & Lim, C. S. (2020). Solubility-weighted index: Fast and accurate prediction of protein solubility. *Bioinformatics*, *36*(18), 4691–4698.

Blanc, M., David, F., Abrami, L., Migliozzi, D., Armand, F., Bürgi, J., & van der Goot, F. G. (2015). SwissPalm: Protein palmitoylation database. *F1000Research*, *4*, 261.

Blanc, M., David, F. P. A., & van der Goot, F. G. (2019). SwissPalm 2: Protein S-palmitoylation database. In *Methods in molecular biology* (pp. 203–214). Springer.

Bobrie, A., Colombo, M., Raposo, G., & Théry, C. (2011). Exosome secretion: Molecular mechanisms and roles in immune responses. *Traffic (Copenhagen, Denmark)*, *12*(12), 1659–1668.

Borges, F. T., Reis, L., & Schor, N. (2013). Extracellular vesicles: Structure, function, and potential clinical uses in renal diseases. *Brazilian Journal of Medical and Biological Research*, *46*(10), 824–830.

Buschow, S. I., Liefhebber, J. M., Wubbolts, R., & Stoorvogel, W. (2005). Exosomes contain ubiquitinated proteins. *Blood Cells, Molecules, and Diseases*, *35*, 398–403.

Camino, T., Lago-Baameiro, N., Bravo, S. B., Molares-Vila, A., Sueiro, A., Couto, I., Baltar, J., Casanueva, E. F., & Pardo, M. (2021). Human obese white adipose tissue sheds depot-specific extracellular vesicles and reveals candidate biomarkers for monitoring obesity and its comorbidities. *Translational Research*, *2*, 85–102.

Campanella, C., D'Anneo, A., Gammazza, A. M., Bavisotto, C. C., Barone, R., Emanuele, S., Cascio, F. L., Mocciaro, E., Fais, S., Macario, E. C. D., Macario, A. J., Cappello, F., & Lauricella, M. (2015). The histone deacetylase inhibitor SAHA induces HSP60 nitration and its extracellular release by exosomal vesicles in human lung-derived carcinoma cells. *Oncotarget*, *7*, 28849–28867.

Carnino, J. M., Ni, K., & Jin, Y. (2020). Post-translational modification regulates formation and cargo-loading of extracellular vesicles. *Frontiers in Immunology*, *11*, 948.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423.

Coumans, F. A., Brisson, A. R., Buzas, E. I., Dignat-George, F., Drees, E. E., El-Andaloussi, S., Emanueli, C., Gasecka, A., Hendrix, A., Hill, A. F., Lacroix, R., Lee, Y., van Leeuwen, T. G., Mackman, N., Mäger, I., Nolan, J. P., van der Pol, E., Pegtel, D. M., Sahoo, S., … Nieuwland, R. (2017). Methodological guidelines to study extracellular vesicles. *Circulation Research*, *120*(10), 1632–1648.

da Silva, J. M., Santiago, V. F., Rosa-Fernandes, L., Marinho, C. R., & Palmisano, G. (2021). Protein glycosylation in extracellular vesicles: Structural characterization and biological functions. *Molecular Immunology*, *135*, 226–246.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., & Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Research*, *34*(2), W362–W365.

de Groot, N., Pallares, I., Aviles, F. X., Vendrell, J., & Ventura, S. (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Structural Biology*, *5*(1), 18.

Dhondt, B., Geeurickx, E., Tulkens, J., Van Deun, J., Vergauwen, G., Lippens, L., Miinalainen, I., Rappu, P., Heino, J., Ost, P., Lumen, N., De Wever, O., & Hendrix, A. (2020). Unravelling the proteomic landscape of extracellular vesicles in prostate cancer by density-based fractionation of urine. *Journal of Extracellular Vesicles*, *9*(1), 1736935.

Dozio, V., & Sanchez, J.-C. (2017). Characterisation of extracellular vesicle-subsets derived from brain endothelial cells and analysis of their protein cargo modulation after tnf exposure. *Journal of Extracellular Vesicles*, *6*(1), 1302705.

Flemming, J. P., Hill, B. L., Haque, M. W., Raad, J., Bonder, C. S., Harshyne, L. A., Rodeck, U., Luginbuhl, A., Wahl, J. K., Tsai, K. Y., Wermuth, P. J., Overmiller, A. M., & Mahoney, M. G. (2020). miRNA- and cytokine-associated extracellular vesicles mediate squamous cell carcinomas. *Journal of Extracellular Vesicles*, *9*, 1790159.

Gámez-Valero, A., Beyer, K., & Borràs, F. E. (2019). Extracellular vesicles, new actors in the search for biomarkers of dementias. *Neurobiology of Aging*, *74*, 15–20.

Gandham, S., Su, X., Wood, J., Nocera, A. L., Alli, S. C., Milane, L., Zimmerman, A., Amiji, M., & Ivanov, A. R. (2020). Technologies and standardization in research on extracellular vesicles. *Trends in Biotechnology*, *38*, 1066–1098.

Gerlach, J. Q., & Griffin, M. D. (2016). Getting to know the extracellular vesicle glycome. *Molecular BioSystems*, *12*, 1071–1081.

Gonzales, P. A., Pisitkun, T., Hoffert, J. D., Tchapyjnikov, D., Star, R. A., Kleta, R., Wang, N. S., & Knepper, M. A. (2008). Large-scale proteomics and phosphoproteomics of urinary exosomes. *Journal of the American Society of Nephrology*, *20*, 363–379.

Greening, D. W., Gopal, S. K., Xu, R., Simpson, R. J., & Chen, W. (2015). Exosomes and their roles in immune regulation and cancer. *Seminars in Cell & Developmental Biology*, *40*, 72–81.

Guruprasad, K., Reddy, B. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, *4*(2), 155–161.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2014). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*, *43*, D512–D520.

Huang, H., Arighi, C. N., Ross, K. E., Ren, J., Li, G., Chen, S.-C., Wang, Q., Cowart, J., Vijay-Shanker, K., & Wu, C. H. (2017). iPTMnet: An integrated resource for protein post-translational modification network discovery. *Nucleic Acids Research*, *46*, D542–D550.

Huang-Doran, I., Zhang, C.-Y., & Vidal-Puig, A. (2017). Extracellular vesicles: Novel mediators of cell communication in metabolic disease. *Trends in Endocrinology & Metabolism*, *28*(1), 3–18.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., De Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., & Sigrist, C. J. (2007). The 20 years of prosite. *Nucleic Acids Research*, *36*(Suppl 1), D245–D249.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589.

Kalra, H., Drummen, G. P., & Mathivanan, S. (2016). Focus on extracellular vesicles: Introducing the next small big thing. *International Journal of Molecular Sciences*, *17*(2), 170.

Karimi, N., Cvjetkovic, A., Jang, S. C., Crescitelli, R., Feizi, M. A. H., Nieuwland, R., Lötvall, J., & Lässer, C. (2018). Detailed analysis of the plasma extracellular vesicle proteome after separation from lipoproteins. *Cellular and Molecular Life Sciences*, *75*(15), 2873–2886.

Keerthikumar, S., Chisanga, D., Ariyaratne, D., Al Saffar, H., Anand, S., Zhao, K., Samuel, M., Pathan, M., Jois, M., Chilamkurti, N., Gangoda, L., & Mathivanan, S. (2016). Exocarta: A web-based compendium of exosomal cargo. *Journal of Molecular Biology*, *428*(4), 688–692.

Kim, D.-K., Kang, B., Kim, O. Y., sic Choi, D., Lee, J., Kim, S. R., Go, G., Yoon, Y. J., Kim, J. H., Jang, S. C., Park, K.-S., Choi, E.-J., Kim, K. P., Desiderio, D. M., Kim, Y.-K., Lötvall, J., Hwang, D., & Gho, Y. S. (2013). EVpedia: An integrated database of high-throughput data for systemic analyses of extracellular vesicles. *Journal of Extracellular Vesicles*, *2*, 20384.

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M. R., & Cebrat, S. (2007). The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms. *BMC Genomics [Electronic Resource]*, *8*, 6.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., & Petersen, B. (2019). Netsurfp- 2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, *87*, 520–527.

Klont, F., Bras, L., Wolters, J. C., Ongay, S., Bischoff, R., Halmos, G. B., & Horvatovich, P. (2018). Assessment of sample preparation bias in mass spectrometrybased proteomics. *Analytical Chemistry*, *90*, 5405–5413.

Koriyama, Y., & Furukawa, A. (2018). S-nitrosylation regulates cell survival and death in the central nervous system. *Neurochemical Research*, *43*, 50–58.

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580.

Kumar, R., & Dhanda, S. K. (2020). Bird eye view of protein subcellular localization prediction. *Life*, *10*(12), 347.

Kurotani, A., Tokmakov, A. A., Sato, K.-I., Stefanov, V. E., Yamada, Y., & Sakurai, T. (2019). Localization-specific distributions of protein pI in human proteome are governed by local pH and membrane charge. *BMC Molecular and Cell Biology*, *20*(1), 36.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132.

Lautenbacher, L., Samaras, P., Muller, J., Grafberger, A., Shraideh, M., Rank, J., Fuchs, S. T., Schmidt, T. K., The, M., Dallago, C., Wittges, H., Rost, B., Krcmar, H., Kuster, B., & Wilhelm, M. (2021). ProteomicsDB: Toward a FAIR open-source resource for life-science research. *Nucleic Acids Research*, *50*(D1), D1541–D1552.

Linares, R., Tan, S., Gounou, C., Arraud, N., & Brisson, A. R. (2015). High-speed centrifugation induces aggregation of extracellular vesicles. *Journal of Extracellular Vesicles*, *4*, 29509.

Liu, B., Leng, L., Sun, X., Wang, Y., Ma, J., & Zhu, Y. (2020). Ecmpride: Prediction of human extracellular matrix proteins based on the ideal dataset using hybrid features with domain evidence. *PeerJ*, *8*, e9066.

Lobry, J., & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 escherichia coli chromosome-encoded genes. *Nucleic Acids Research*, *22*(15), 3174–3180.

Lötvall, J., Hill, A. F., Hochberg, F., Buzás, E. I., Vizio, D. D., Gardiner, C., Gho, Y. S., Kurochkin, I. V., Mathivanan, S., Quesenberry, P., Sahoo, S., Tahara, H., Wauben, M. H., Witwer, K. W., & Théry, C. (2014). Minimal experimental requirements for definition of extracellular vesicles and their functions: A position statement from the international society for extracellular vesicles. *Journal of Extracellular Vesicles*, *3*, 26913.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 00–00.

Mariscal, J., Vagner, T., Kim, M., Zhou, B., Chin, A., Zandian, M., Freeman, M. R., You, S., Zijlstra, A., Yang, W., & Vizio, D. D. (2020). Comprehensive palmitoyl-proteomic analysis identifies distinct protein signatures for large and small cancer-derived extracellular vesicles. *Journal of Extracellular Vesicles*, *9*, 1764192.

Martinéz-Greene, J. A., Hernández-Ortega, K., Quiroz-Baez, R., Resendis- Antonio, O., Pichardo-Casas, I., Sinclair, D. A., Budnik, B., Hidalgo-Miranda, A., Uribe-Querol, E., del Pilar Ramos-Godínez, M., & Martínez-Martínez, E. (2021). Quantitative proteomic analysis of extracellular vesicle subgroups isolated by an optimized method combining polymer-based precipitation and size exclusion chromatography. *Journal of Extracellular Vesicles*, *10*(6), e12087.

Mathieu, M., Martin-Jaular, L., Lavieu, G., & Thery, C. (2019). Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication. *Nature Cell Biology*, *21*, 9–17.

Miranda, K. C., Bond, D. T., McKee, M., Skog, J., Paunescu, T. G., Da Silva, N., Brown, D., & Russo, L. M. (2010). Nucleic acids within urinary exosomes/microvesicles are potential biomarkers for renal disease. *Kidney International*, *78*(2), 191–199.

Moreno-Gonzalo, O., Villarroya-Beltri, C., & Sanchez-Madrid, F. (2014). Posttranslational modifications of exosomal proteins. *Frontiers in Immunology*, *5*, 383.

Paolicelli, R. C., Bergamini, G., & Rajendran, L. (2019). Cell-to-cell communication by extracellular vesicles: Focus on microglia. *Neuroscience*, *405*, 148–157.

Parolini, I., Federici, C., Raggi, C., Lugini, L., Palleschi, S., Milito, A. D., Coscia, C., Iessi, E., Logozzi, M., Molinari, A., Colone, M., Tatti, M., Sargiacomo, M., & Fais, S. (2009). Microenvironmental pH is a key factor for exosome traffic in tumor cells. *Journal of Biological Chemistry*, *284*, 34211–34222.

Pathan, M., Fonseka, P., Chitti, S. V., Kang, T., Sanwlani, R., Deun, J. V., Hendrix, A., & Mathivanan, S. (2018). Vesiclepedia 2019: A compendium of RNA, proteins, lipids and metabolites in extracellular vesicles. *Nucleic Acids Research*, *47*, D516–D519.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Picciotto, S., Romancino, D. P., Buffa, V., Cusimano, A., Bongiovanni, A., & Adamo, G. (2020). Post-translational lipidation in extracellular vesicles: Chemical mechanisms, biological functions and applications. In *Advances in biomembranes and lipid self-assembly* (pp. 83–111). Elsevier.

Ras-Carmona, A., Gomez-Perosanz, M., & Reche, P. A. (2021). Prediction of unconventional protein secretion by exosomes. *BMC Bioinformatics [Electronic Resource]*, *22*(1), 333.

Robbins, P. D., & Morelli, A. E. (2014). Regulation of immune responses by extracellular vesicles. *Nature Reviews Immunology*, *14*, 195–208.

Romancino, D. P., Buffa, V., Caruso, S., Ferrara, I., Raccosta, S., Notaro, A., Campos, Y., Noto, R., Martorana, V., Cupane, A., Giallongo, A., d'Azzo, A., Manno, M., & Bongiovanni, A. (2018). Palmitoylation is a post-translational modification of Alix regulating the membrane organization of exosome-like small extracellular vesicles. *Biochimica Et Biophysica Acta (BBA)—General Subjects*, *1862*, 2879–2887.

Samaras, P., Schmidt, T., Frejno, M., Gessulat, S., Reinecke, M., Jarzab, A., Zecha, J., Mergner, J., Giansanti, P., Ehrlich, H.-C., Aiche, S., Rank, J., Kienegger, H., Krcmar, H., Kuster, B., & Wilhelm, M. (2019). ProteomicsDB: A multi-omics and multiorganism resource for life science research. *Nucleic Acids Research*, *48*, D1153–D1163.

Shah, S. M. A., & Ou, Y.-Y. (2021). TRP-BERT: Discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT. *Computers in Biology and Medicine*, *137*, 104821.

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., & Bucher, P. (2002). Prosite: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, *3*(3), 265–274.

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2012). New and continuing developments at prosite. *Nucleic Acids Research*, *41*(D1), D344–D347.

Simpson, R. J., Kalra, H., & Mathivanan, S. (2012). Exocarta as a resource for exosomal research. *Journal of Extracellular Vesicles*, *1*(1), 18374.

Smith, V. L., Jackson, L., & Schorey, J. S. (2015). Ubiquitination as a mechanism to transport soluble mycobacterial and eukaryotic proteins to exosomes. *The Journal of Immunology*, *195*, 2722–2730.

Sodar, B. W., Kittel, Á., Paloczi, K., Vukman, K. V., Osteikoetxea, X., Szabo-Taylor, K., Nemeth, A., Sperlagh, B., Baranyai, T., Giricz, Z., Wiener, Z., Turiák, L., Drahos, L., Pállinger, É., Vékey, K., Ferdinandy, P., Falus, A., & Buzás, E. I. (2016). Low-density lipoprotein mimics blood plasma-derived exosomes and microvesicles during isolation and detection. *Scientific Reports*, *6*(1), 1–12.

Stam, J., Bartel, S., Bischoff, R., & Wolters, J. C. (2021). Isolation of extracellular vesicles with combined enrichment methods. *Journal of Chromatography B*, *1169*, 122604.

Stomberski, C. T., Hess, D. T., & Stamler, J. S. (2019). Protein s-nitrosylation: Determinants of specificity and enzymatic regulation of s-nitrosothiol-based signaling. *Antioxidants & Redox Signaling*, *30*, 1331–1351.

Théry, C., Witwer, K. W., Aikawa, E., Alcaraz, M. J., Anderson, J. D., Andriantsitohaina, R., Antoniou, A., Arab, T., Archer, F., Atkin-Smith, G. K., Ayre, D. C., Bach, J. M., Bachurski, D., Baharvand, H., Balaj, L., Baldacchino, S., Bauer, N. N., Baxter, A. A., Bebawy, M., … Zuba-Surma, E. K. (2018). Minimal information for studies of extracellular vesicles 2018 (misev2018): A position statement of the international society for extracellular vesicles and update of the misev2014 guidelines. *Journal of Extracellular Vesicles*, *7*(1), 1535750.

Tian, Y., Gong, M., Hu, Y., Liu, H., Zhang, W., Zhang, M., Hu, X., Aubert, D., Zhu, S., Wu, L., & Yan, X. (2019). Quality and efficiency assessment of six extracellular vesicle isolation methods by nano-flow cytometry. *Journal of Extracellular Vesicles*, *9*, 1697028.

UniProt Consortium (2019). Uniprot: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515.

van Gils, J. H. M., Gogishvili, D., van Eck, J., Bouwmeester, R., van Dijk, E., & Abeln, S. (2022). How sticky are our proteins? Quantifying hydrophobicity of the human proteome. *Bioinformatics Advances*, *2*, vbac002.

van Niel, G., D'Angelo, G., & Raposo, G. (2018). Shedding light on the cell biology of extracellular vesicles. *Nature Reviews Molecular Cell Biology*, *19*, 213–228.

Veerman, R. E., Teeuwen, L., Czarnewski, P., Akpinar, G. G., Sandberg, A., Cao, X., Pernemalm, M., Orre, L. M., Gabrielsson, S., & Eldh, M. (2021). Molecular evaluation of five different isolation methods for extracellular vesicles reveals different clinical applicability and subcellular origin. *Journal of Extracellular Vesicles*, *10*(9), e12128.

Walker, J. D., Maier, C. L., & Pober, J. S. (2009). Cytomegalovirus-infected human endothelial cells can stimulate allogeneic CD4+memory t cells by releasing antigenic exosomes. *The Journal of Immunology*, *182*, 1548–1559.

Wang, D., Liang, Y., & Xu, D. (2018). Capsule network for protein post-translational modification site prediction. *Bioinformatics*, *35*, 2386–2394.

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J., & Xu, D. (2020). MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, *48*, W140–W146.

Wang, Y., Zhou, Z., Leylek, T., Tan, H., Sun, Y., Parkinson, F., & Wang, J.-F. (2015). Protein cysteine s-nitrosylation inhibits vesicular uptake of neurotransmitters. *Neuroscience*, *311*, 374–381.

Wang, Y., Zhou, Z., Tan, H., Zhu, S., Wang, Y., Sun, Y., Li, X.-M., & Wang, J.-F. (2016). Nitrosylation of vesicular transporters in brain of amyloid precursor protein/presenilin 1 double transgenic mice. *Journal of Alzheimer's Disease*, *55*, 1683–1692.

Watson, L. S., Hamlett, E. D., Stone, T. D., & Sims-Robinson, C. (2019). Neuronally derived extracellular vesicles: An emerging tool for understanding Alzheimer's disease. *Molecular Neurodegeneration*, *14*, 6.

Waury, K., Willemse, E. A. J., Vanmechelen, E., Zetterberg, H., Teunissen, C. E., & Abeln, S. (2022). Bioinformatics tools and data resources for assay development of fluid protein biomarkers. *Biomarker Research*, *10*(1), 83.

Welton, J. L., Webber, J. P., Botos, L.-A., Jones, M., & Clayton, A. (2015). Ready-made chromatography columns for extracellular vesicle isolation from plasma. *Journal of Extracellular Vesicles*, *4*(1), 27269.

Whiteside, T. L. (2015). The potential of tumor-derived exosomes for noninvasive cancer monitoring. *Expert Review of Molecular Diagnostics*, *15*(10), 1293–1310.

Witwer, K. W., Buzás, E. I., Bemis, L. T., Bora, A., Lässer, C., Lötvall, J., Hoen, E. N. N., Piper, M. G., Sivaraman, S., Skog, J., Théry, C., Wauben, M. H., & Hochberg, F. (2013). Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *Journal of Extracellular Vesicles*, *2*, 20360.

Wolfers, J., Lozier, A., Raposo, G., Regnault, A., Théry, C., Masurier, C., Flament, C., Pouzieux, S., Faure, F., Tursz, T., Angevin, E., Amigorena, S., & Zitvogel, L. (2001). Tumor-derived exosomes are a source of shared tumor rejection antigens for CTL cross-priming. *Nature Medicine*, *7*, 297–303.

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene set knowledge discovery with enrichr. *Current Protocols*, *1*(3), e90.

Yáñez-Mó, M., Siljander, P. R.-M., Andreu, Z., Bedina Zavec, A., Borràs, F. E., Buzas, E. I., Buzas, K., Casal, E., Cappello, F., Carvalho, J., Colás, E., Cordeiro-da Silva, A., Fais, S., Falcon-Perez, J. M., Ghobrial, I. M., Giebel, B., Gimona, M., Graner, M., Gursel, I., … De Wever, O. (2015). Biological properties of extracellular vesicles and their physiological functions. *Journal of Extracellular Vesicles*, *4*(1), 27066.

Zhang, X., Borg, E. G. F., Liaci, A. M., Vos, H. R., & Stoorvogel, W. (2020). A novel three step protocol to isolate extracellular vesicles from plasma or cell culture medium with both high yield and purity. *Journal of Extracellular Vesicles*, *9*, 1791450.

Zhao, L., Poschmann, G., Waldera-Lupa, D., Rafiee, N., Kollmann, M., & Stuhler, K. (2019). Outcyte: A novel tool for predicting unconventional protein secretion. *Scientific Reports*, *9*(1), 1–9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Waury, K., Gogishvili, D., Nieuwland, R., Chatterjee, M., Teunissen, C. E., & Abeln, S. (2024). Proteome encoded determinants of protein sorting into extracellular vesicles. *Journal of Extracellular Biology*, *3*, e120. https://doi.org/10.1002/jex2.120