Mechanistic Prediction of Extracellular Vesicle Surface-Accessible Proteins from Whole Extracellular Vesicle Proteomics

Patrick Mertens
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
Caretronic d.o.o.
Kranj, Slovenia

Anton Gradišek
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Mitja Luštrek
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

Proteomics of whole extracellular vesicles (EVs) identifies vesicle-associated proteins but does not directly determine which proteins are externally accessible on intact vesicles. We evaluated whether protein-level structural and mechanistically motivated descriptors can prioritize surface-accessible proteins. Whole-vesicle proteomics from Kirkemo et al. identified 1,546 unique proteins. A separate experiment was used to determine which of these proteins were detectable on the vesicle surface. Four proteins could not be assigned reliably to either class, so they were excluded from model training. This left 540 proteins detected in the surface experiment and 1,002 proteins not detected in the surface experiment. We combined 123 baseline structural/mechanistic descriptors with 43 EV-oriented descriptors covering membrane architecture, trafficking motifs, topology/curvature, and disorder. Four classifiers were evaluated by stratified five-fold out-of-fold cross-validation. The combined mechanistic representation reached a pooled ROC-AUC of 0.77 and average precision of 0.65 with Extra Trees; the positive prevalence, and therefore the no-skill baseline for average precision, was 0.35. Database-derived annotations describing the cellular compartments in which proteins are known or expected to occur increased ROC-AUC to approximately 0.88, but were analyzed separately because they provide prior biological knowledge about localization. These results support mechanistic protein descriptors as a useful prioritization layer for experimental EV-surface studies rather than a substitute for direct surface measurements.

Keywords

extracellular vesicles, surface proteomics, protein localization, machine learning, bioinformatics

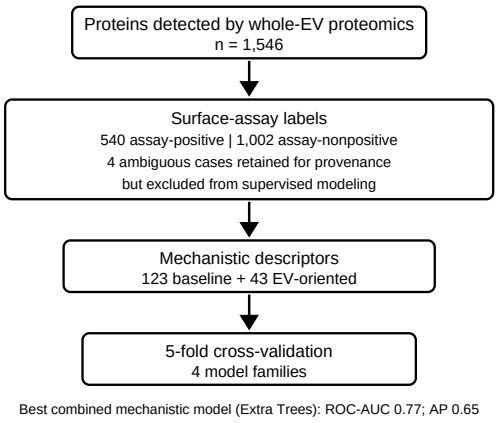


Figure 1: Analysis workflow. Surface labels are defined only among proteins detected by whole-EV proteomics; localization annotations are evaluated separately from the primary mechanistic representation.

1 Introduction

Extracellular vesicles (EVs) are membrane-bound particles involved in intercellular communication and comprise heterogeneous populations generated through distinct biogenic routes [14, 7]. Their externally accessible proteins are particularly relevant to vesicle–cell interactions, affinity capture, and biomarker or targeting strategies. However, proteins detected in whole-EV proteomics include membrane, luminal, cytosolic, and co-isolated proteins; detection in an EV preparation therefore does not establish external surface accessibility.

Kirkemo et al. developed WGA–HRP surface proteomics for cells and small EVs and released the underlying mass spectrometry data as PXD028523 [8]. We use these data to ask a computationally distinct question: among proteins already detected in whole-EV proteomics, can structural and mechanistic descriptors prioritize proteins supported by an experimental surface assay? We treat this as a ranking/classification problem. Database-derived annotations describing the cellular compartments associated with proteins are evaluated separately because they encode prior biological knowledge from existing experimental and computational evidence, rather than properties derived directly from the protein itself. Such annotations may also be incomplete, unavailable, or inapplicable to newly characterized or engineered proteins. Their performance therefore provides an annotation-informed reference rather than evidence of de novo mechanistic prediction. Figure 1 summarizes the design.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.124

2 Materials and Methods

2.1 Proteins Detected by Whole-EV Proteomics and Surface Labels

The prediction set was constructed from proteins detected in the whole-EV mass-spectrometry experiment after removing standard quality-control entries representing decoys and contaminants. Surface labels were derived from the corresponding EV surface-proteomics experiment of Kirkemo et al. [8]. A protein was labeled assay-positive when it was detected at least once in any of eight surface-assay measurements (four Control and four Myc runs); operationally, detection was defined as a spectral count greater than zero.

The resulting analysis set contained 1,546 representative proteins. Four cases in which surface evidence mapped to a different member of the same MaxQuant protein group were retained as ambiguous for provenance but left unlabeled rather than transferring evidence across paralogs, and were therefore excluded from supervised modeling. The supervised cohort therefore contained 1,542 proteins: 540 assay-positive and 1,002 assay-nonpositive. An assay-nonpositive label is an operational non-detection state and is not interpreted as proof that a protein is biologically internal.

2.2 Mechanistic and Annotation Representations

Each protein was represented by a numerical feature vector suitable for input to the machine-learning classifiers. A protein can be described by its amino-acid sequence—the ordered chain of amino acids from which it is built—together with available information about its structure and organization.

The baseline representation contained 123 structural and mechanistic descriptors. These variables summarize general protein properties such as amino-acid composition and physicochemical characteristics, membrane-spanning regions, signal peptides, glycosylation and other modification sites, and known protein domains or short functional sequence patterns. The descriptors were calculated from the amino-acid sequence together with curated structural, membrane-topology, modification, and protein-domain information from biological databases [13, 2]. The baseline representation therefore combines properties calculated directly from the protein sequence with previously curated structural information.

We added 43 EV-oriented descriptors intended to represent biological mechanisms that could influence whether a protein becomes accessible on the surface of an extracellular vesicle. These descriptors were organized into five groups.

The first group describes membrane-related properties. These variables summarize features such as the number and geometry of membrane-spanning regions and properties of the surrounding amino acids associated with different membrane environments [12].

The second and third groups describe intracellular protein sorting and trafficking. Cells contain molecular machinery that recognizes short patterns in a protein's amino-acid sequence and uses them as signals for directing proteins through particular cellular pathways. Some of the added descriptors count patterns associated with endosomal sorting and vesicle formation, including motifs associated with the ESCRT sorting machinery [6], while others represent motifs involved in endocytic trafficking between cellular compartments [3]. Computationally, these are numerical features; for example, a descriptor can represent the number of times a particular sorting pattern occurs within an appropriate region of a protein.

The fourth group describes membrane topology and curvature-related properties. For membrane-associated proteins, topology describes which regions are annotated as cytosolic and which are non-cytosolic or lumen-facing. The corresponding descriptors summarize properties such as charge asymmetry between these regions, compactness of non-cytosolic/lumen-facing regions, and an N-terminal helical hydrophobic moment. The hydrophobic moment measures how hydrophobic amino acids are distributed around a hypothetical helical structure and was calculated using Kyte–Doolittle hydropathy values with the Eisenberg vector formulation [1, 9, 4].

The fifth group describes protein disorder and local sequence complexity. These descriptors quantify regions whose amino-acid composition suggests a reduced tendency to form a stable compact structure, using FoldIndex-based measures and local Shannon-entropy measures [10, 15]. When topology information explicitly indicated that a region was absent, the corresponding feature was represented as zero; when the topology itself was unresolved, the value was retained as missing. Two exploratory trafficking variables and additional redundant, topology-availability, and quality-control/provenance fields were excluded from the 43-feature primary representation.

Together, the 123 baseline descriptors and 43 EV-oriented descriptors formed the 166-variable primary mechanistic representation.

For comparison, we separately evaluated 41 database-derived annotation variables describing the cellular compartments associated with proteins and related localization information. These annotations encode prior biological knowledge from existing experimental and computational studies rather than being derived primarily from the protein's own measurable properties. Because such information may be absent or incomplete, particularly for newly characterized or engineered proteins, the annotation variables were evaluated as a separate reference setting and were not included in the primary mechanistic model.

2.3 Model Evaluation

We evaluated elastic-net logistic regression ($C = 1$, ℓ_1 ratio = 0.5), Random Forest, Extra Trees [5], and histogram gradient boosting.

Random Forest and Extra Trees used 500 trees with square-root feature sampling; histogram gradient boosting used learning rate 0.05, 300 iterations, and L2 regularization 0.01.

The primary combined mechanistic setting contained 166 predictors (123 baseline + 43 EV-oriented).

Stratified five-fold cross-validation with shuffling and random seed 42 generated exactly one out-of-fold (OOF) score per protein and model. Median imputation was fitted within each training fold; scaling for elastic-net logistic regression was likewise fold-local. To reduce bias toward the more frequent assay-nonpositive class, training examples were weighted inversely to class frequency so that the two classes contributed approximately equally during model fitting.

The primary discrimination metrics were pooled-OOF ROC-AUC and average precision (AP); AP is informative in class-imbalanced binary classification and its no-skill baseline equals the positive prevalence [11]. To characterize variability across the five cross-validation folds, ROC-AUC and AP were also calculated separately for each held-out fold and summarized as mean

Table 1: Five-fold cross-validation performance, reported as mean ± standard deviation across the held-out folds. p is the number of predictors. EN, elastic-net logistic regression; RF, Random Forest; ET, Extra Trees; HGB, histogram gradient boosting.

Setting	p	Model	ROC-AUC	AP
Baseline	123	EN	0.745 ± 0.034	0.603 ± 0.060
		RF	0.763 ± 0.026	0.634 ± 0.048
		ET	0.769 ± 0.030	0.640 ± 0.053
		HGB	0.760 ± 0.030	0.647 ± 0.049
EV-oriented	43	EN	0.679 ± 0.052	0.560 ± 0.070
		RF	0.698 ± 0.038	0.562 ± 0.049
		ET	0.687 ± 0.042	0.565 ± 0.055
		HGB	0.692 ± 0.047	0.555 ± 0.048
Combined	166	EN	0.736 ± 0.039	0.610 ± 0.062
		RF	0.770 ± 0.024	0.644 ± 0.041
		ET	0.772 ± 0.026	0.660 ± 0.047
		HGB	0.757 ± 0.031	0.641 ± 0.052
Annotation	41	EN	0.870 ± 0.013	0.748 ± 0.038
		RF	0.879 ± 0.018	0.749 ± 0.029
		ET	0.867 ± 0.016	0.727 ± 0.027
		HGB	0.874 ± 0.012	0.729 ± 0.035
Combined+ann.	207	EN	0.855 ± 0.021	0.742 ± 0.059
		RF	0.882 ± 0.023	0.771 ± 0.049
		ET	0.884 ± 0.022	0.775 ± 0.044
		HGB	0.874 ± 0.017	0.751 ± 0.040

± standard deviation. Comparisons between feature representations were descriptive; no formal hypothesis tests of differences in ROC-AUC or AP were performed.

After cross-validation, each model was also refitted using all 1,542 labeled proteins to generate candidate-ranking scores and descriptive feature importances. Because these scores were produced by models trained on the same proteins, they were not used to estimate predictive performance.

3 Results

Table 1 summarizes the main OOF results.

The 43 EV-oriented descriptors alone showed moderate discrimination: ROC-AUC ranged from 0.68 to 0.70 across the four model families, with AP from 0.54 to 0.56.

The 123-feature baseline representation was stronger (ROC-AUC 0.74–0.77; AP 0.60–0.64).

The combined 166-feature mechanistic representation performed best among the mechanistic settings: Extra Trees reached ROC-AUC 0.77 and AP 0.65.

Because the no-skill baseline for AP equals the positive prevalence, AP=0.65 should be interpreted relative to the assay-positive prevalence of 0.35.

Using the combined 166-feature mechanistic model, the 1,542 labeled proteins were ranked by their Extra Trees out-of-fold prediction scores. Of the 50 highest-ranked proteins, 38 were assay-positive, corresponding to a precision-at-50 (P@50) of 0.76.

The incremental effect of the EV-oriented descriptors was modest. Within Extra Trees, adding the 43 EV-oriented descriptors to the 123-feature baseline increased pooled OOF ROC-AUC from 0.768 to 0.771 ($\Delta = 0.004$) and AP from 0.627 to 0.646 ($\Delta = 0.019$). These differences are reported descriptively; no formal hypothesis test between feature sets was performed. Accordingly, we interpret the EV-oriented descriptors as complementary information rather than as evidence of a statistically established improvement over the baseline.

Database-derived localization annotations were substantially more predictive. Models using the 41 annotation variables reached ROC-AUC 0.87–0.88, while combining the mechanistic and annotation variables reached ROC-AUC 0.88 and AP 0.77 at best.

This higher performance is consistent with the annotation variables already containing prior biological knowledge about the cellular compartments in which proteins are known or expected to occur. In the Random Forest model fitted to all 1,542 labeled proteins, the most influential annotation variables included plasma-membrane Gene Ontology terms and UniProt membrane-related keywords and location information. However, database annotations may be incomplete or unavailable, particularly for poorly characterized or engineered proteins. We therefore report annotation-informed performance separately from the primary mechanistic model rather than interpreting the approximately 0.88 ROC-AUC as performance derived from protein properties alone.

Additional trafficking-related variables outside the primary 43-feature EV-oriented set were also tested. Their inclusion had little effect: for Extra Trees, both the primary 166-feature setting and the setting containing these additional variables round to ROC-AUC 0.77 and AP 0.65. They were therefore excluded from the primary mechanistic representation.

Descriptive feature importances from the Extra Trees model fitted to all 1,542 labeled proteins were distributed across several types of protein information. Important variables represented short sequence patterns, protein-domain information, molecular size, membrane-related physicochemical properties, and measures of protein disorder and sequence complexity. This indicates that the fitted model did not rely on a single type of protein information. These feature importances describe associations used by the model and are not interpreted as causal biological mechanisms.

4 Discussion and Conclusion

The cross-validation results show that structural and mechanistically motivated protein descriptors contain predictive information about assay-defined EV surface accessibility.

The combined mechanistic representation achieved ROC-AUC 0.77 and AP 0.65 without using database-derived localization annotations. This supports these protein descriptors as a computational prioritization layer for proteins identified by whole-EV proteomics.

The stronger performance of models using database-derived localization annotations is informative rather than contradictory. These annotations contain prior biological knowledge about the cellular compartments in which proteins are known or expected to occur. Therefore, the approximately 0.88 ROC-AUC should be interpreted as performance when such prior localization information is available, rather than as prediction from structural and mechanistically motivated protein properties alone.

The primary mechanistic model addresses the more challenging question of whether proteins supported by the experimental surface assay can be prioritized without using this prior localization information.

Several limitations constrain interpretation.

First, the experimental labels are assay-derived proxies. Non-detection does not establish true absence from the EV surface,

and measured surface accessibility may depend on EV subtype, membrane orientation, protein abundance, and experimental sensitivity.

Second, the data originate from a specific RWPE-1/Myc small-EV system [8]. Kirkemo et al. studied EVs derived from control and Myc-induced RWPE-1 prostate epithelial cells, so generalization to other cell types, EV populations, and biofluids remains untested.

Third, model feature importances indicate which variables contributed to predictions made by the fitted model, but they do not establish that those protein properties cause EV surface accessibility. Importance can also be distributed among correlated predictors. Consequently, these values can help identify potentially informative features, but they cannot by themselves establish the biological mechanisms responsible for surface localization.

Fourth, the predictive evaluation was internal: each protein was evaluated in a held-out fold during five-fold cross-validation, but no independent dataset was used for external validation. Validation on independent surface-proteomics datasets and targeted experimental testing of high-ranking candidates are therefore required before broader biological generalization.

Within these limits, the results provide a reproducible computational framework for prioritizing candidate surface-accessible proteins among proteins detected by whole extracellular-vesicle proteomics.

References

[1] Bruno Antonny. 2011. Mechanisms of membrane curvature sensing. *Annual Review of Biochemistry*, 80, 101–123. doi:10.1146/annurev-biochem-052809-155121.

[2] Matthias Blum et al. 2025. Interpro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53, D1, D444–D456. doi:10.1093/nar/gkae1082.

[3] Juan S. Bonifacino and Linton M. Traub. 2003. Signals for sorting of transmembrane proteins to endosomes and lysosomes. *Annual Review of Biochemistry*, 72, 395–447. doi:10.1146/annurev.biochem.72.121801.161800.

[4] David Eisenberg, Robert M. Weiss, and Thomas C. Terwilliger. 1982. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299, 371–374. doi:10.1038/299371a0.

[5] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63, 1, 3–42. doi:10.1007/s10994-006-6226-1.

[6] James H. Hurley and Phyllis I. Hanson. 2010. Membrane budding and scission by the ESCRT machinery: it's all in the neck. *Nature Reviews Molecular Cell Biology*, 11, 556–566. doi:10.1038/nrm2937.

[7] Raghu Kalluri and Valerie S. LeBleu. 2020. The biology, function, and biomedical applications of exosomes. *Science*, 367, 6478, eaau6977. doi:10.1126/science.aau6977.

[8] Lisa L. Kirkemo, Susanna K. Elledge, Jiuling Yang, James R. Byrnes, Jeff E. Glasgow, Robert Blelloch, and James A. Wells. 2022. Cell-surface tethered promiscuous biotinylators enable comparative small-scale surface proteomic analysis of human extracellular vesicles and cells. *eLife*, 11, e73982. doi:10.7554/eLife.73982.

[9] Jack Kyte and Russell F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157, 1, 105–132. doi:10.1016/0022-2836(82)90515-0.

[10] Jaime Prilusky, Clifford E. Felder, Tzviya Zeev-Ben-Mordehai, Edwin H. Rydberg, Orna Man, Jacques S. Beckmann, Israel Silman, and Joel L. Sussman. 2005. Foldindex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21, 16, 3435–3438. doi:10.1093/bioinformatics/bti537.

[11] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10, 3, e0118432. doi:10.1371/journal.pone.0118432.

[12] Hayley J. Sharpe, Tim J. Stevens, and Sean Munro. 2010. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell*, 142, 1, 158–169. doi:10.1016/j.cell.2010.05.037.

[13] The UniProt Consortium. 2025. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53, D1, D609–D617. doi:10.1093/nar/gkae1010.

[14] Guillaume van Niel, Gisela D'Angelo, and Graça Raposo. 2018. Shedding light on the cell biology of extracellular vesicles. *Nature Reviews Molecular Cell Biology*, 19, 213–228. doi:10.1038/nrm.2017.125.

[15] John C. Wootton and Scott Federhen. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17, 2, 149–163. doi:10.1016/0097-8485(93)85006-X.