MLMarker: A machine learning framework for tissue inference and biomarker discovery

Tine Claeys[1,2], Sam van Puyenbroeck[1,2], Kris Gevaert[1,2], Lennart Martens[1,2,3]*

*Correspondence should be addressed to lennart.martens@ugent.be

1 VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

2 Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

3 BioOrganic Mass Spectrometry Laboratory (LSMBO), IPHC UMR 7178, University of Strasbourg, CNRS, Strasbourg, 67000, France

Abstract

Mass spectrometry-based proteomics enables high-throughput profiling of protein expression across tissues but interpreting complex or sparse datasets remains challenging. We here present MLMarker, a machine learning-based tool trained on healthy human tissue proteomic data to compute continuous tissue similarity scores for new datasets. Using a Random Forest model trained across 34 tissue types, MLMarker predicts probabilistic tissue scores and explains these predictions using SHAP values at the single protein level for biological interpretability. To address missingness, a penalty factor can be used to correct for absent proteins contributing to predictions.

We demonstrate MLMarker's utility across three public datasets: (i) cerebral melanoma metastases, where brain-like proteomic profiles were linked to poor treatment response and invasive phenotypes, revealing 241 differentially expressed proteins; (ii) a pan-cancer FFPE dataset, where tissue origin predictions reached 79% accuracy using broad tissue mappings despite the absence of cancer-specific training data; and (iii) cerebrospinal fluid and plasma, where the model identified brain and pituitary origin in biofluid samples.

MLMarker offers an explainable, robust framework for hypothesis generation in proteomics, applicable across tumours, biofluids, and multi-tissue studies. It is available as a Python package and Streamlit GUI at https://mlmarker.streamlit.app.

Keywords

Tissue prediction, Mass spectrometry-based proteomics, Machine learning, Public data reuse, AI

Background

Mass spectrometry (MS)-based proteomics has become a cornerstone for protein expression analysis in biological systems. Recent technological advancements in sample preparation[1–3], instrumentation[4–6], and computational methods[7–10] have drastically increased sensitivity, throughput and coverage, enabling the detection and quantification of thousands of proteins across diverse biological samples, all while accommodating ever decreasing amounts of sample. However, as the scale and speed of proteomic dataset generation keeps growing, drawing meaningful biological conclusions remains a challenge due to the inherent complexity and vastness of the proteome[11,12].

Many downstream interpretation methods rely primarily on the detection of differentially expressed proteins, and these approaches often fall short of providing a holistic, multi-layered comparative analysis. Traditional approaches, such as pairwise differential expression analysis (DEA), aim to narrow the scope of data interpretation by focusing on predefined group comparisons. While these methods offer statistical rigor, they can limit the exploration of more complex, non-obvious relationships within the data. Additionally, they can be significantly influenced by preprocessing steps such as quantification methods, normalization, and missing value imputation, resulting in a combinatorial explosion of possible outcomes, in turn leading to inconsistent results[13,14]. Capturing the full differential landscape of a sample while considering its broader context, thus remains a significant challenge. The field therefore stands to gain substantially from the development of innovative approaches that can more comprehensively address the data's complexity.

Interestingly, the large amounts of data being deposited into ProteomeXchange[15] repositories such as PRIDE[16] and MassIVE[16] provide the potential to improve these analysis methods. Indeed, this treasure trove of proteomics data has already led to the development of essential community knowledgebases such as PeptideAtlas[17], ProteomicsDB[18], and Scop3P[19]. Furthermore, it has enabled technical advances such as the development of machine learning models like MS2PIP[20] and DeepLC[21] that allow greatly improved MS data interpretation through rescoring approaches[7,22]. While the reuse of public data for biological discovery has so far remained limited, it has demonstrated potential in identifying and validating biomarkers[23], and in conducting meta-analyses on post-translational modifications[24,25].

Here, we introduce another way of making use of this vast public proteomics data with MLMarker, a machine learning (ML) based tool based on our previously developed tissue prediction model[26]. MLMarker is designed to complement proteome analysis by using its ML model to compare each sample to a reference atlas of healthy human tissue proteomes. Unlike conventional approaches that rely on rigid, pre-assigned group comparisons that typically all originate in the same study, MLMarker provides a continuous, probabilistic measure of tissue similarity across multiple tissues

simultaneously. This enables a more nuanced view of sample classification, revealing unexpected tissue relationships that may indicate disease adaptations and/or uncover novel biological phenomena. Importantly, MLMarker thus shifts tissue prediction from a simple classification task to a powerful framework for hypothesis generation, as samples with unexpected tissue similarities can point to new biological mechanisms or disease processes [27–29].

We demonstrate the model's applicability on three public datasets: (i) a reanalysis of cerebral melanoma metastases from Zila *et al*. (PXD007592)[30], (ii) a reanalysis of cerebrospinal fluid from Yamana *et al*. (PXD008029)[31], and (iii) a reanalysis of the pan-cancer FFPE tissue dataset from Tüshaus *et al*. (MSV000095036)[32]. Together these three case studies illustrate the versatility of MLMarker, showing its potential not only in understanding cancer metastases, but also in analysing biofluid samples and mapping tissue identities in complex multi-cancer datasets. By leveraging a tissue-centric approach, MLMarker offers a new perspective on proteomics data, complementing existing methods and expanding the scope of biological discovery.

MLMarker is available as a pip installable python package and a graphical user interface through the streamlit framework at https://mlmarker.streamlit.app.

Results

Reprocessed data from cerebral melanoma (PXD007592)

Zila *et al*. analysed proteomic profiles of cerebral melanoma metastases to grasp the difference between good and poor treatment response to MAP kinase inhibitors (MAPKi). The samples were obtained through surgical excision, and response state to MAPKi treatment was determined through progression-free survival (PFS). Good responders had a PFS ≥ 6 months and low disease progression, while poor responders had a PFS < 3 months and fast disease progression. The authors observed small molecular differences between good and poor responders, with poor responders exhibiting an epithelial-mesenchymal transition (EMT) signature and drug resistance mechanisms, while good responders showed higher immune activation. To note is that this different response to treatment was determined on a phenotypic level and not on the proteotype of the tumour. Ultimately, the authors associated good response patients with proliferative melanoma tumours and poor response patients with invasive melanoma tumours, but, at 1% false discovery rate (FDR) and a Perseus s_0 of 0.5, could only identify fourteen differentially expressed proteins between the two response states[30].

This dataset was selected to explore whether cerebral melanoma metastases are classified by MLMarker as brain, reflecting their site of growth, or as skin, reflecting their tissue of origin. As metastatic tumours may retain features of both origins, they were expected to provide a useful model for testing if and how MLMarker resolves their origin. As mentioned, the original analysis identified only minor molecular differences between good and poor responders, with limited separation in unsupervised clustering, suggesting that relevant biological structures remained hidden. By applying MLMarker, we aimed to determine whether tissue similarity scores could uncover this latent structure and offer complementary insight beyond treatment response.

Reprocessing of the dataset was performed using ionbot[33] for peptide identification, FlashLFQ[34] for quantification, and MSqRob2[35] for differential expression analysis. Twenty proteins were found significantly differentially expressed (adjusted p-value < 0.01, Fig 1a) and this low variance between response states is also reflected in the UMAP clustering of protein expression values which shows no clear separation between response states (Fig 1b). Only one protein was indicated as differentially expressed in both workflows: the eukaryotic translation initiation factor 3 subunit I (See Supplementary Table S1).

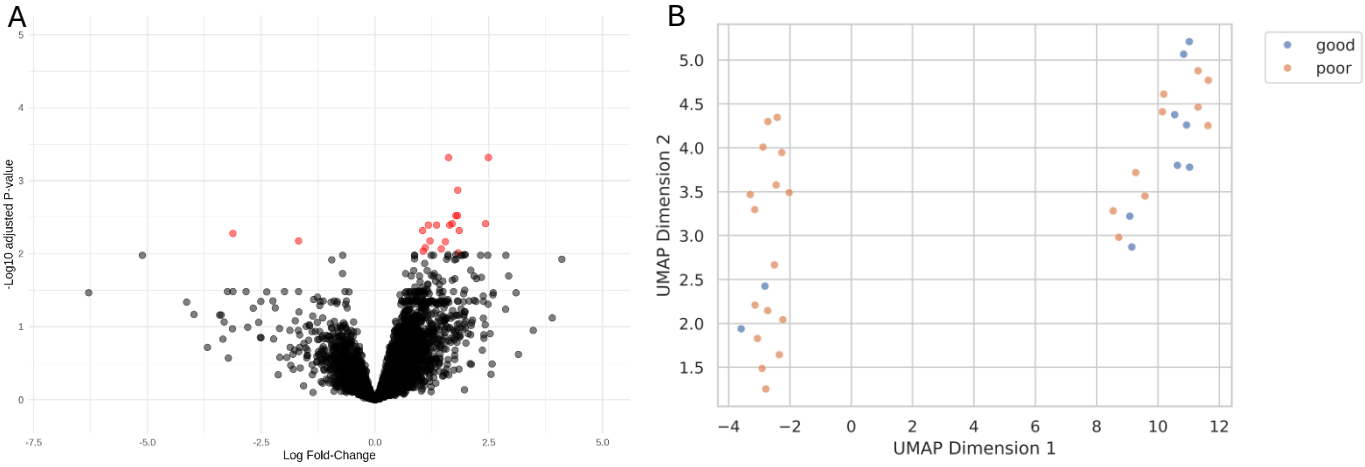


Fig 1 a) Volcano plot displaying differential expression analysis between good and poor responders. Log fold change on x-axis, and -log$_{10}$ adjusted p-value after Bonferroni correction on y-axis. Twenty proteins with adjusted p-value < 0.01 are coloured red. b) UMAP of protein expression between good and poor responders.

Applying MLMarker to the protein expression data revealed a striking difference. Poor responders had significantly higher brain similarity scores than good responders (Mann-Whitney U test, p = 0.035, Fig 2a). However, brain similarity does not perfectly correlate with the response state as can be seen in the UMAP clustering on prediction level, in which two distinct clusters can be observed, correlating with the brain similarity metric rather than response state (Fig 2b). This resulted in four subpopulations: poor responders with high brain similarity (twenty samples) and with low brain similarity (six samples), and good responders with high brain similarity (four samples) and low similarity (six samples). Notably, when samples were not predicted as brain, their highest-scoring alternative tissues were either lung or monocytes.

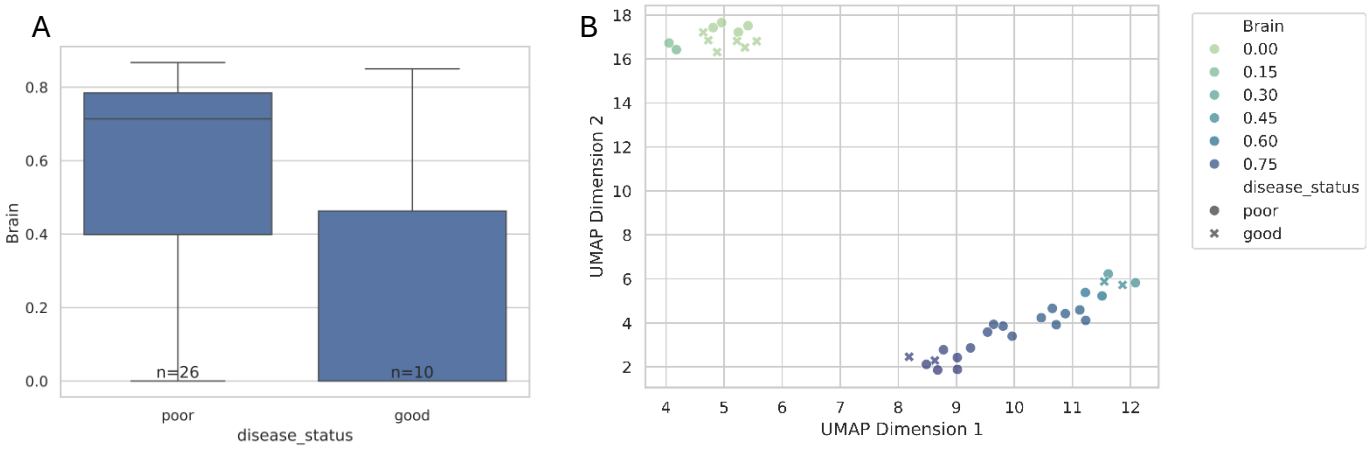


Fig 2a) Distribution of brain similarity values as output by MLMarker, per response state. The two distributions are significantly different (Mann-Whitney U test, p = 0.035). 2b) UMAP clustering of MLMarker output, coloured by brain similarity and dot shape indicating response state.

To determine whether MLMarker's predictions were influenced by data quality, we analysed (i) sample content, and (ii) overlap with the MLMarker fixed feature set. Indeed,

MLMarker relies on a set 5,979 of predefined proteins seen during training, meaning it can only process and generate predictions using the proteins in this feature set. Any missing proteins are assigned a value of zero, potentially affecting the accuracy and reliability of the predictions. No trend was observed between the number of identified features, dynamic range, and brain prediction scores (Fig 3), suggesting that MLMarker's classification is not driven by data quality or quantity artifacts.

Fig 3. Top panel: number of MLMarker features per sample, coloured according to brain prediction. Bottom panel: summed spectral abundance factor per sample, coloured according to brain prediction. No clear correlation between sample content and prediction is observed.

To further investigate the tumour profiles, we took two distinct approaches. First, we utilized the insights from the MLMarker model by interpreting Shapley Additive exPlanations (SHAP) values, which offered a deeper understanding of the model's decision-making process. A negative SHAP value indicates that a protein negatively influences the predicted tissue classification, while a positive SHAP value suggests a positive impact. It is also important to emphasize that the absence of proteins can be informative in this context, and the non-detection of proteins can thus influence the prediction process.

Second, we applied the tissue classifications generated by MLMarker to define groups for DEA with MSqRob2[35], comparing brain against non-brain samples. This step

complements the original differential expression analysis, which differentiated between good and poor responders, providing additional insight into the proteomic differences between these groups.

SHAP analysis of MLMarker decision

To understand the model's decision process, we plotted the SHAP values across the four subpopulations (Fig 4). These values quantify the contribution of each feature to the model's predictions. Notably, brain-predicted samples (good_brain and poor_brain) exhibit higher SHAP values, suggesting that several proteins positively influence the model's decision-making. In contrast, non-brain-predicted samples (good_notbrain and poor_notbrain) show lower SHAP values, indicating that the model relies less on tissue-specific markers. Instead, classification appears driven by a broader proteomic profile, potentially incorporating less tissue-specific features.

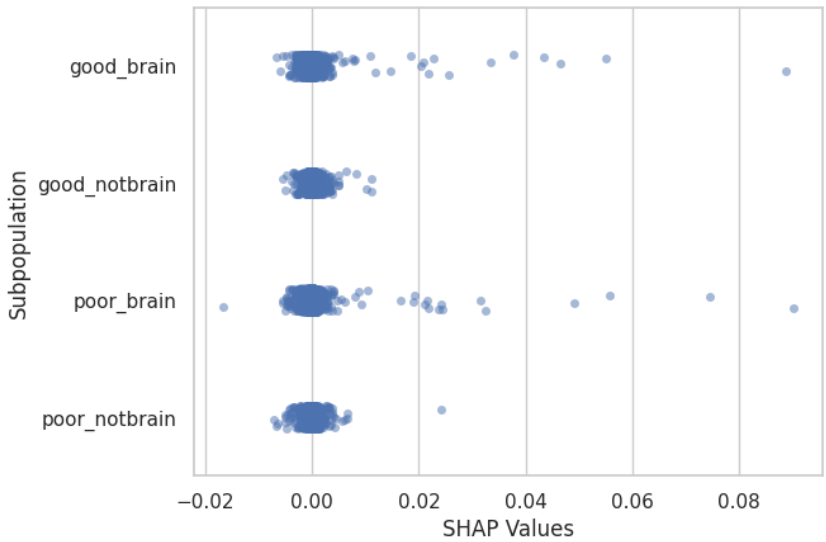


Fig 4. SHAP value distribution of subpopulations showing brain predicted populations with a wider value distribution demonstrating the presence of highly tissue specific proteins, which are lacking in the non-brain predicted population.

To gain a more in-depth understanding of decision-driving proteins in each subpopulation, we selected the top 1% highest-ranking SHAP value proteins per subpopulation for a gene ontology (GO) enrichment, revealing a clear tissue-specific signature (Fig S1). Unsurprisingly, brain-predicted samples, regardless of response state were enriched for brain-associated processes. In contrast, non-brain predicted samples showed a sharp decline in brain-related functions among poor responders and a complete absence in good responders. Instead, good responders exhibit an increase in immune-related processes, which are entirely absent in the other groups.

Differential expression analysis with MSqRob

To fully demonstrate MLMarker's potential as a hypothesis generator, we performed two types of differential expression analysis (DEA). First, we compared responders and non-

responders, replicating the original study's approach. Our results identified twenty differentially expressed proteins (DEPs) with an adjusted p-value < 0.01, a similar number of DEPs compared to the original results. However, when we performed the same analysis between brain-predicted and non-brain-predicted samples, we observed a high increase in DEPs, totalling 241 proteins. Of these, 207 were upregulated within brain-predicted samples, and 34 were upregulated within not brain-predicted samples. These results far exceed the number of DEPs identified in the original good vs. poor responder comparison (Fig 5, top panel).

Next, we performed differential expression analysis (DEA) between good and poor responders within the brain-predicted and non-brain-predicted subpopulations separately. This analysis revealed no differentially expressed proteins (DEPs) in brain-predicted samples, while 25 DEPs were identified in non-brain-predicted samples, suggesting that the MAPK inhibitor response is only weakly influenced by changes in overall protein expression (Fig 5, bottom panel).

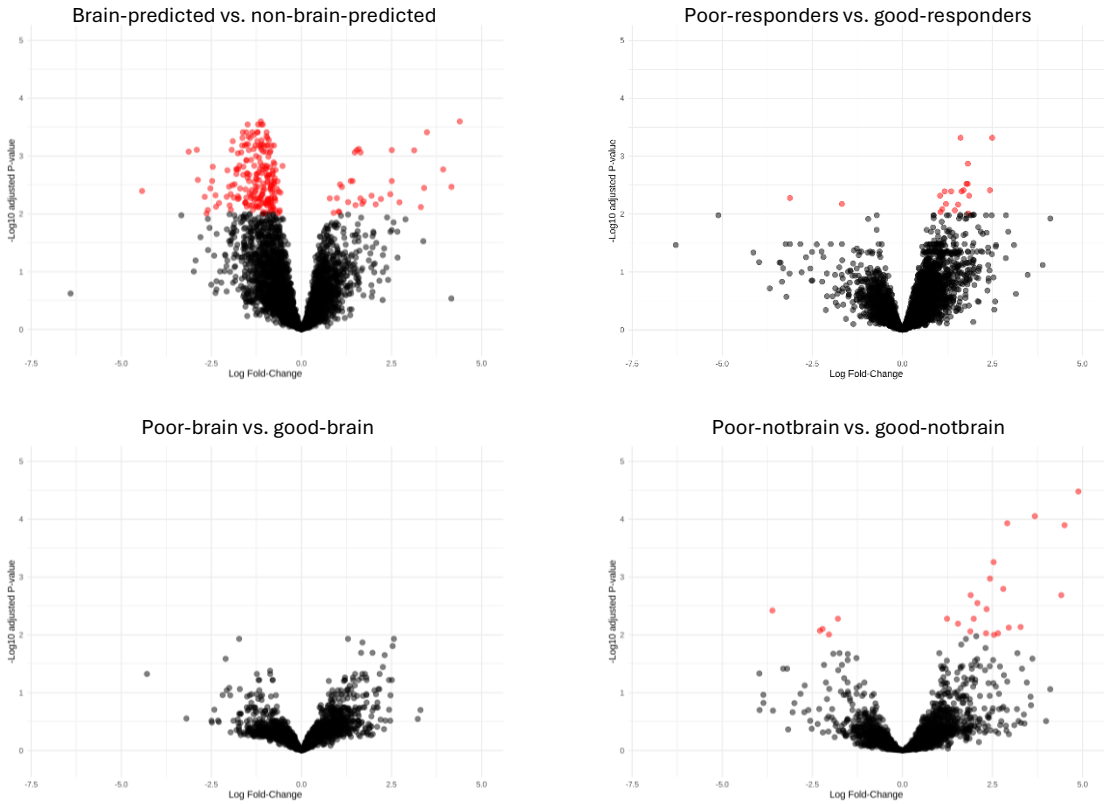


Fig 5: Differential expression analysis volcano plots with log fold change on X-axis, and -log_{10} adjusted p-value following Bonferroni correction on Y-axis. Red dots indicate proteins with adjusted p-value >= 0.01.Top panel: Comparison of brain-predicted vs. non-brain-predicted samples (left) identified 241 DEPs (adjusted p-value < 0.01), a tenfold increase over the responder vs. non-responder comparison (right).Bottom panel: Within brain-predicted (left) and non-brain-predicted (right) subpopulations DEA

between good and poor responders, no DEPs were identified in brain-predicted samples, while 25 DEPs were found in non-brain-predicted samples.

Further analysis of DEPs in brain vs non-brain

Analysing the 241 differentially expressed proteins (DEPs; Table S1), a striking observation is their limited brain specificity. Indeed, GO enrichment analysis reveals two distinct molecular signatures (Fig S2. Proteins overexpressed in brain-predicted samples include some that are involved in neuronal processes, yet their functions extend beyond the brain. Instead, they are associated with metabolism, extracellular space, secretion, protein binding, and enzyme activity. Additionally, these proteins are linked to metabolic pathways such as the TCA cycle, and Parkinson's disease. Notably, the molecular functions "cadherin binding" and "heat shock protein binding" emerge as enriched, both of which were previously associated with an invasive tumour phenotype[30]. In contrast, the 31 proteins overexpressed in non-brain-predicted samples are predominantly linked to protein activation cascades, blood coagulation, and extracellular processes.

From a published list of 76 known protein markers of tumour invasion[36], ten were found among the DEPs in brain-predicted samples, while none were present in the non-brain predicted DEPs. These ten proteins are primarily involved in metabolism, cytoskeleton organization and signalling (Table 1).

Identified marker	UniProt id	Full name	Function	Tissue specifcity (UniProt)	Log2 fold change (not_brain – brain)
1433E	P62258	14-3-3 protein epsilon	signaling protein	low	-1.30
1433Z	P63104	14-3-3 protein zeta/delta	signaling protein	low	-1.44
ACLY	P53396	ATP-citrate synthase	metabolism, cholesterol and fatty acid	low	-1.07
COF1	P23528	Cofilin-1	actin binding, cytoskeletal organization, neural tube	low	-1.18
FAS	P49327	Fatty acid synthase	synthesis of long-chain saturated fatty acids	adipose tissue, breast	-1.53
FSCN1	Q16658	Fascin	actin binding, cell protrusions	low	-1.72
HS90A	P07900	Heat shock protein HSP 90-alpha	chaperone for cell cycle control and signal transduction proteins	low	-1.23
PHB	P35232	Prohibitin	chaperone for signaling, mitochondrion and transcription factors in the nucleus	low	-1.25
STMN1	P16949	Stathmin	microtubule filament system	brain, lymphoid tissue	-1.22
TPIS	P60174	Triosephosphate isomerase	metabolism glycolysis and gluconeogenesis	skeletal muscle	-1.36

Table 1: Differentially expressed proteins that are known markers of tumour invasion according to Pouliquen et al.[36]

Circularity and double dipping

To evaluate the potential for circular analysis, often referred to as "double dipping", we examined the correlation between SHAP values from MLMarker predictions and adjusted p-values from MSqRob. Circular analysis occurs when the same data is used both to select and to test a hypothesis, leading to inflated statistical significance.

First, we examined the proportion of DEPs in the MLMarker features, followed by a correlation analysis between SHAP values from MLMarker predictions and adjusted p-values from MSqRob. Comparing the p-value distributions between MLMarker features and not MLMarker features revealed a similar pattern (Fig. 6). This similarity persisted when restricting to DEPs with $\log_2$fold changes greater than |1|. Within the MLMarker features the proportion of DEPs (13.6 %) closely matches that within not MLMarker features (11.9%). A Mann-Whitney U test showed no significant difference in the p-value distribution for all proteins but indicated a slight skew in more differentially regulated proteins among MLMarker features when considering $\log_2$ fold change >|1| (p = 0.04).

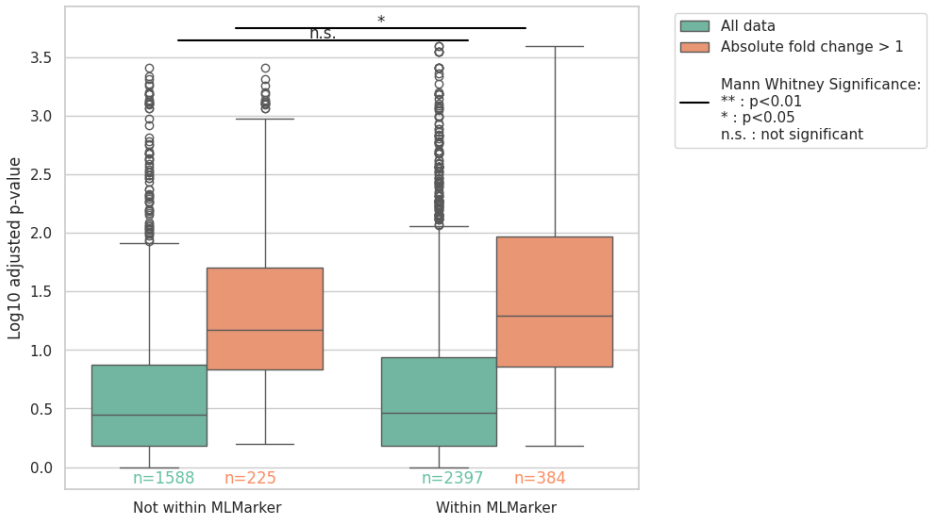


Fig 6. Boxplot showing the distribution of $-\log_{10}$ adjusted p-values for proteins outside (not within) and inside (within) the MLMarker feature space. All proteins, independent of their $\log_2$ fold change, are shown in green, while proteins with absolute $\log_2$fold change > 1 are highlighted in orange. Mann–Whitney U tests indicate no significant difference between groups for all proteins (p > 0.05), while proteins with absolute fold change > 1 show a significant difference (p = 0.04).

Next, we evaluated whether the p-value of MLMarker features correlated with the models SHAP values (Fig. 7). Plotting maximal absolute SHAP values against the minimal $\log_{10}$-adjusted p-values from MSqRob, we observed no significant correlation: Spearman ρ = 0.004 (p = 0.938). This indicates that while SHAP-positive proteins may show enriched differential signal as a group as seen in the previous analysis, there is no systematic scaling relationship between feature importance and statistical significance. Overall, the

limited correlation indicates partial overlap between SHAP values and differential expression results without evidence of circular analysis.

This aligns with the distinct nature of the measures: SHAP values capture tissue-specific patterns learned from external data, while differential expression reflects cohort-specific statistics. Thus, the results support SHAP-based interpretation as a complementary approach rather than one prone to circular reasoning or data leakage.

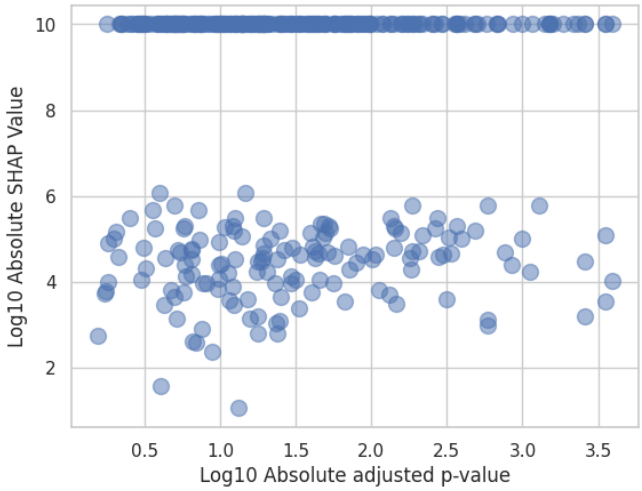


Fig 7. Scatterplot comparing on the Y-axis the log_{10}-transformed maximal absolute SHAP value and on the X-axis the minimal log-transformed adjusted p-values. Each point represents a protein feature. No significant correlation between SHAP values and adjusted p-values was identified (Spearman $\rho = 0.004$, $p = 0.938$).

Reprocessed data from cerebrospinal fluid (PXD008029)

In this study (PXD008029[37]), an experimental workflow was developed to maximize protein identifications from human cerebrospinal fluid (CSF) and human plasma. These biofluids underwent depletion of highly abundant proteins and fractionation to enhance the detection of lower-abundance proteins, including potential tissue-leakage markers. We suspected that this sample type posed a challenge for MLMarker as MLMarker was not trained on any biofluids and relies on the previously discussed fixed feature space. Indeed, MLMarker can only use the proteins it was trained upon and requires all proteins to be associated with a quantitative value. To mitigate skewed predictions in sparse datasets, two strategies can be employed. First, the relationship between prediction certainty and protein content can be assessed. Second, a sample with all features containing an intensity of zero can serve as a baseline reference. The latter approach is built into the penalty factor of MLMarker, which preserves SHAP values for detected proteins while penalizing missing proteins that strongly contribute to the prediction. This method, rather than relying on abstract loss functions or complicated models, ensures biological relevance and preserves the interpretability of the model. The penalty factor subtracts the SHAP values of missing proteins with their corresponding SHAP value in a fully zeroed-out baseline sample, ensuring that missing proteins do not disproportionately influence tissue classification, while still reflecting the strong independent contributions of detected proteins. This approach is recommended for samples that deviate from the MLMarker training samples such as biofluids, cell lines, and organoids, or for comparatively sparse samples such as single cells. The impact of the penalty factor is evident when predicting an empty sample: without correction, the model assigns a 19% probability to bone marrow, which drops to 0.5% upon applying the penalty factor (Fig S3).

When predicting the CSF samples, pituitary gland and brain are the two highest ranking tissues independent of using a penalty factor. However, introducing the penalty factor reduces the probability of lower ranking tissue classes such as bone marrow, testis, lung and monocytes. The plasma predictions demonstrate a similar drop in bone marrow. In those samples, adipose tissue is the first ranking tissue followed by monocytes. This can be explained by the lack of any blood-related tissue within the MLMarker training data (Fig 8).

The balance between quantity and prediction confidence is evident in the CSF samples. A strong correlation exists between the number of identified spectra (1% FDR) and the probability of predicting pituitary gland and brain (Pearson correlation of respectively 0.60 and 0.45). However, even low-protein-content samples can yield high-confidence predictions if they contain highly specific proteins (see Fig S4). For the plasma samples, the correlation is less pronounced for adipose tissue and monocytes (Pearson correlation of respectively 0.23 and 0.44) which can again be explained by the complete absence of any blood-related training data in MLMarker.

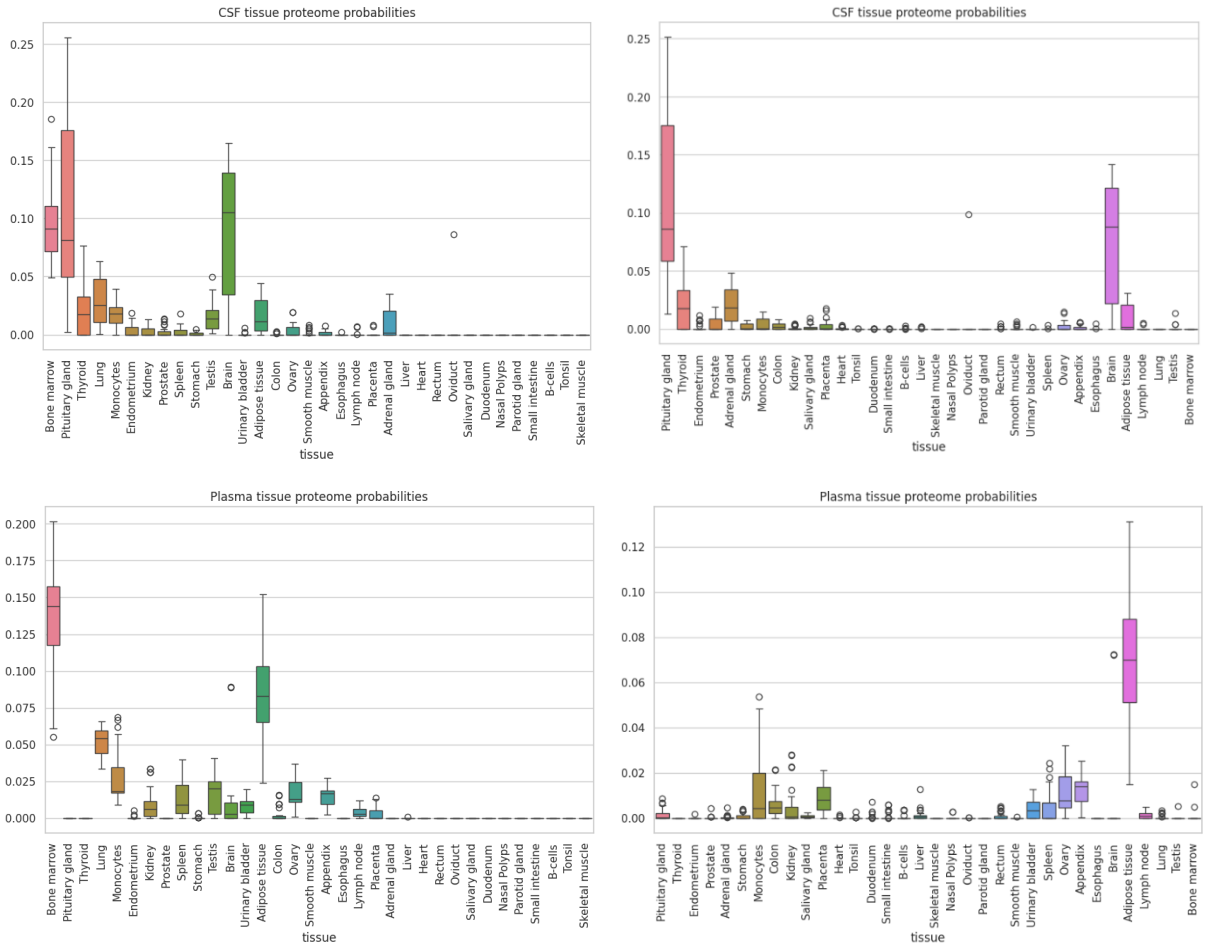


Fig 8. Top panels: CSF tissue probabilities before (left) and after (right) penalty factor correction. Bottom panels: Plasma tissue probabilities before (left) and after (right) penalty factor correction.

SHAP analysis of CSF samples within the Streamlit application

MLMarker offers an intuitive and user-friendly interface through a Streamlit application, which is available online and requires no installation. This allows researchers without coding expertise to easily explore the tissue predictions of their samples, making advanced proteomic data analysis accessible to a broader user group. What follows is an example of how MLMarker's functionalities can be used to gain deeper insights into the proteins driving tissue classifications, facilitated by the interactive features of the Streamlit interface.

The SHAP analysis functionality is visualized through tissue-specific bar plots, which highlight both positively and negatively contributing proteins. For example, in a CSF sample with high predicted probabilities for both the pituitary gland and brain, SHAP values clearly reveal an abundance of proteins reinforcing these tissues, while also showing a substantial number of proteins negatively influencing the bone marrow prediction (Fig 9a).

Further enhancing the analysis, MLMarker offers an interactive tissue-specific visualization, where bar size reflects the magnitude of each protein's contribution. This feature allows users to easily browse and identify potentially interesting proteins, enhancing their ability to pinpoint key drivers of tissue classification (Fig 9b). To explore the relationship between protein contributions across tissues, we also generated a scatterplot of adjusted SHAP values for the same CSF sample with high pituitary gland and brain probabilities. This scatterplot reveals how different proteins contribute to each tissue prediction (Fig 9c). (i) those that positively contribute to both tissues, Q1; (ii) those that enhance pituitary gland classification while negatively impacting brain prediction, Q2; (iii) those that strengthen brain classification while reducing pituitary gland probability, Q3 and; (iv) those that negatively contribute to both tissues, Q4.

This interactive analysis, seamlessly integrated into the Streamlit application, thus offers an accessible, user-friendly way for researchers to explore and evaluate protein contributions to tissue classifications, making it easier to identify potential biomarkers.

A closer examination of the top tissue-specific contributors confirms their biological relevance. The twenty most influential proteins in the pituitary gland classification predominantly belong to secretory and neuronal pathways, aligning with the known function of this tissue. Likewise, the key proteins driving brain classification are well-established brain-associated markers, several of which are already known to be present in CSF (Tables 2 and 3).

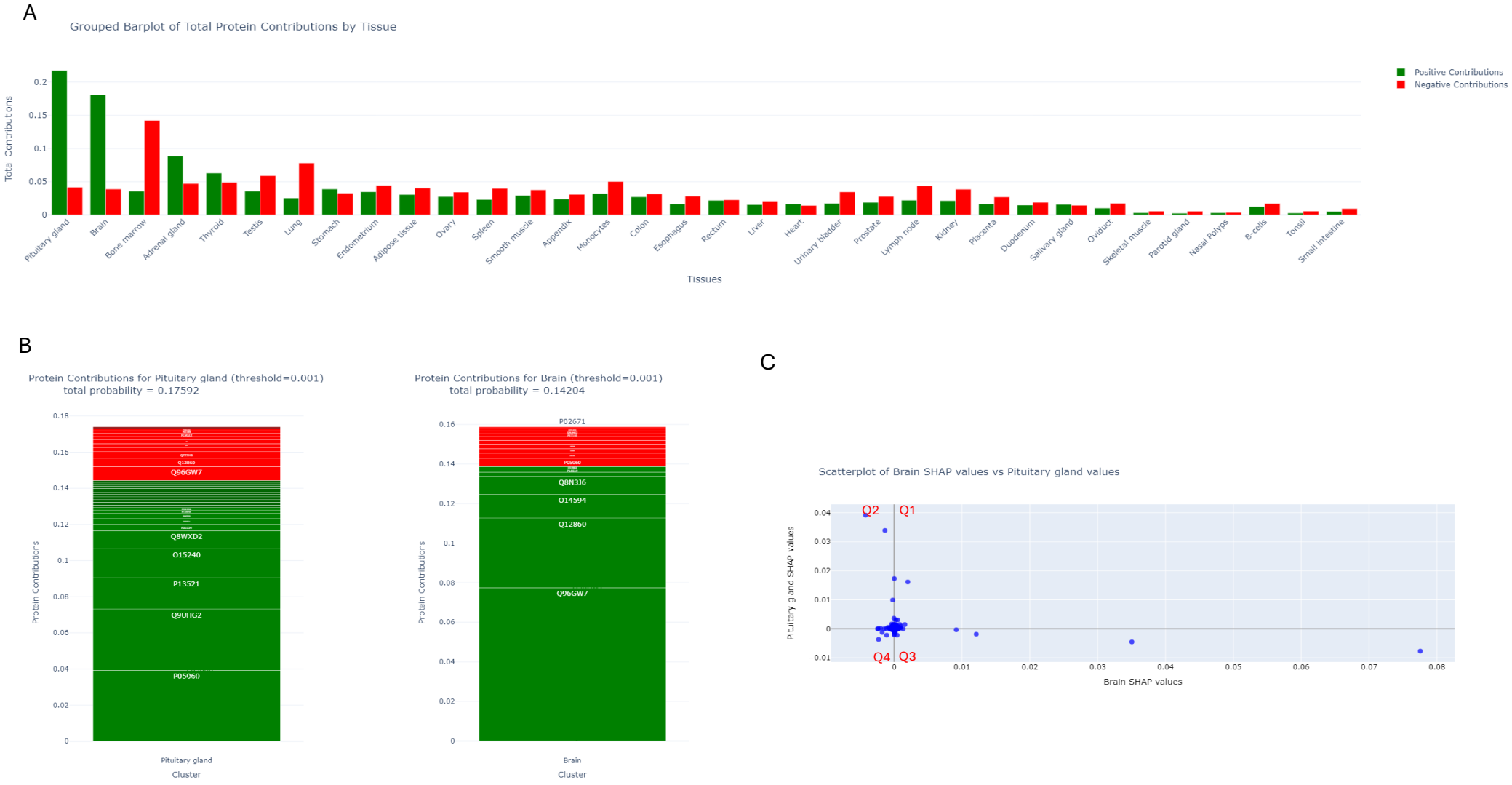


Fig 9 A) Sample level grouped bar plot showing positive (green) and negative (red) SHAP contributions per tissue. B) Tissue level bar plot showing positive and negative SHAP contributions for pituitary gland (left) and brain (right). C) Scatterplot of SHAP protein values within brain (x-axis) and pituitary gland (y-axis) showing the interplay of protein across tissues with quartiles indicated.

PITUITARY GLAND	UniProt ID	SHAP value	protein_names	tissue_specificity	RNA tissue specific nTPM	Tissue expression cluster
1	P05060	0.03916	Secretogranin-1	CSF and urine	adrenal gland: 3361.4;brain: 677.1;pituitary gland: 1387.3	Cluster 49: Pituitary gland - Hormone signaling
2	Q9UHG2	0.0339	ProSAAS	Brain and pancreas	brain: 776.1;pituitary gland: 1211.0	Cluster 33: Brain - Neuronal signaling
3	P13521	0.01732	Secretogranin-2	CSF and urine	adrenal gland: 349.2;brain: 368.9;pituitary gland: 261.8	Cluster 33: Brain - Neuronal signaling
4	O15240	0.01612	Neurosecretory protein VGF [Cleaved into: Neuroendocrine regulatory peptide-1	Central and peripheral nervous systems	brain: 75.2;pituitary gland: 51.7	Cluster 33: Brain - Neuronal signaling
5	Q8WXD2	0.00993	Secretogranin-3	Detected in urine. Expressed in brain, heart, kidney, liver and skeletal muscle. .	brain: 161.3;pituitary gland: 38.3;retina: 55.0	Cluster 17: Neuronal - Synaptic function
6	P01034	0.00361	Cystatin-C	Parotid gland and saliva	brain: 4419.8;choroid plexus: 4010.9	Cluster 60: Choroid plexus - Transmembrane transport
7	P68871	0.00306	Hemoglobin subunit beta	Red blood cells.	bone marrow: 131241.6	Cluster 77: Bone marrow - Cell proliferation
8	Q06520	0.00302	Sulfotransferase 2A1	Liver and adrenal gland	liver: 1043.0	Cluster 83: Liver - Metabolism
9	P10645	0.0016	Chromogranin-A	CSF, urine and brain	parathyroid gland: 6356.0	Cluster 24: Parathyroid gland - Vesicular transport
10	P02452	0.0016	Collagen alpha-1(I) chain	Bones	cervix: 1284.0;gallbladder: 1170.7;ovary: 1172.9	Cluster 62: Connective tissue - ECM organization
11	P16519	0.00147	Neuroendocrine convertase 2	nan	adrenal gland: 27.6;retina: 34.7;thyroid gland: 55.2	Cluster 17: Neuronal - Synaptic function

12	P04114	0.00147	Apolipoprotein B-100	nan	intestine: 296.0;liver: 1155.6	Cluster 83: Liver - Metabolism
13	P07101	0.00145	Tyrosine 3-monooxygenase	Brain and adrenal gland	adrenal gland: 190.6;brain: 130.3	Cluster 49: Pituitary gland - Hormone signaling
14	P05108	0.00143	Cholesterol side-chain cleavage enzyme, mitochondrial	nan	adrenal gland: 978.1	Cluster 26: Adrenal gland - Steroid metabolism
15	P01019	0.00137	Angiotensinogen	Liver and plasma	liver: 1563.6	Cluster 83: Liver - Metabolism
16	P35900	0.00121	Keratin, type I cytoskeletal 20	Intestinal epithelium	intestine: 629.1	Cluster 45: Intestine - Digestion
17	P25815	0.00117	Protein S100-P	Detected in all of the tissues except brain, testis and small intestine	bone marrow: 1393.2;stomach 1: 1692.9;urinary bladder: 1708.0	Cluster 77: Bone marrow - Cell proliferation
18	P06702	0.00117	Protein S100-A9	myeloid cells	bone marrow: 31844.0;esophagus: 29726.6;vagina: 13605.6	Cluster 77: Bone marrow - Cell proliferation
19	Q9ULL0	0.00113	Acrosomal protein KIAA1210	nan	epididymis: 40.8;seminal vesicle: 20.3;testis: 31.8	Cluster 63: Testis - Spermatid development
20	O75891	0.00111	Cytosolic 10-formyltetrahydrofolate dehydrogenase	Highly expressed in liver, pancreas and kidney. .	liver: 358.6	Cluster 69: Liver - Oxidoreductase activity

Table 2. Top 20 proteins positively contributing to pituitary gland tissue prediction, ranked by SHAP value. Gene names, protein names, and tissue specificity annotations were retrieved from UniProt based on the corresponding UniProt identifiers with nan indicating not annotated natively in UniProt. RNA expression levels and tissue-specific expression clusters were obtained from the Human Protein Atlas (HPA). RNA levels are indicated by normalized transcript per million (nTPM) allowing comparison across all HPA samples. Clusters are assigned based on gene to gene distance and shared nearest neighbor clustering to find clusters of genes with similar expression profiles within the graph. Clusters are then manually annotated using overrepresentation analysis including Gene Ontology, KEGG, Reactome, PanglaoDB, TRRUST and HPA classifications.

BRAIN	UniProt ID	SHAP value	Protein Name	UniProt tissue specificty	HPA RNA tissue specific nTPM	HPA Tissue expression cluster
1	Q96GW7	0.07755	Brevican core protein	Expressed in the retina, CSF and urine	brain: 245.5	Cluster 33: Brain - Neuronal signaling
2	Q12860	0.03504	Contactin-1	Strongly expressed in brain and in neuroblastoma and retinoblastoma cell lines. Lower levels of expression in lung, pancreas, kidney and skeletal muscle. .	brain: 125.0	Cluster 64: Neuronal - Transcription
3	O14594	0.01211	Neurocan core protein	Detected in CSF and Brain	brain: 102.1	Cluster 61: Brain - Neuronal signaling
4	Q8N3J6	0.00916	Cell adhesion molecule 2	nan	brain: 50.2;retina: 92.0	Cluster 17: Neuronal - Synaptic function
5	O15240	0.00202	Neurosecretory protein VGF [Cleaved into: Neuroendocrine regulatory peptide-1	Central and peripheral nervous systems	brain: 75.2;pituitary gland: 51.7	Cluster 33: Brain - Neuronal signaling
6	P16519	0.00159	Neuroendocrine convertase 2	nan	adrenal gland: 27.6;retina: 34.7;thyroid gland: 55.2	Cluster 17: Neuronal - Synaptic function
7	Q14982	0.00132	Opioid-binding protein/cell adhesion molecule	nan	brain: 36.6;parathyroid gland: 10.0;retina: 11.0	Cluster 22: Brain - Synaptic signal transduction
8	P54868	0.00099	Hydroxymethylglutaryl-CoA synthase, mitochondrial	High expression in liver. Low expression in colon, kidney, testis and pancreas. No expression in brain	liver: 2359.8	Cluster 83: Liver - Metabolism
9	P19013	0.00096	Keratin, type II cytoskeletal 4	Epithelial	esophagus: 19104.8	Cluster 29: Squamous epithelium - Innate immune response
10	P07101	0.00087	Tyrosine 3-monooxygenase	Brain and adrenal gland	adrenal gland: 190.6;brain: 130.3	Cluster 49: Pituitary gland - Hormone signaling
11	P09228	0.00085	Cystatin-SA	Parotid gland and saliva	salivary gland: 5804.2	Cluster 38: Salivary gland - Salivary secretion
12	P07196	0.00068	Neurofilament light polypeptide	nan	brain: 183.3;retina: 86.8	Cluster 17: Neuronal - Synaptic function
13	O95340	0.00067	Bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthase 2	Cartilage and adrenal gland	adrenal gland: 131.0	Cluster 11: Liver - Plasma proteins

14	P35558	0.00059	Phosphoenolpyruvate carboxykinase, cytosolic [GTP]	Liver, kidney and adipocytes	kidney: 416.9;liver: 664.7	Cluster 1: Liver & Kidney - Metabolism
15	P02760	0.00058	Protein AMBP	Expressed by the liver, secreted in plasma, urine and CSF	liver: 10607.0	Cluster 31: Liver - Plasma proteins
16	P14770	0.00053	Platelet glycoprotein IX	nan	bone marrow: 5.0;lung: 3.1;lymphoid tissue: 8.7	Cluster 81: Lymphoid tissue & Bone marrow - Innate immune response
17	Q8TDL5	0.00053	BPI fold-containing family B member 1	Epithelium of duodenum, trachea and lung	cervix: 546.4;salivary gland: 264.3;stomach 1: 499.1	Cluster 75: Epithelium - Extracellular exosomes
18	P12036	0.00052	Neurofilament heavy polypeptide	nan	brain: 95.6;prostate: 237.7	Cluster 64: Neuronal - Transcription
19	O15230	0.00051	Laminin subunit alpha-5	Heart, lung, kidney, skeletal muscle, pancreas, retina and placenta		Cluster 66: Non-specific - Transcription
20	Q06520	0.00051	Sulfotransferase 2A1	Liver and adrenal gland	liver: 1043.0	Cluster 83: Liver - Metabolism

Table 3. Top 20 proteins positively contributing to brain tissue prediction, ranked by SHAP value. Gene names, protein names, and tissue specificity annotations were retrieved from UniProt based on the corresponding UniProt identifiers with nan indicating not annotated natively in UniProt. RNA expression levels and tissue-specific expression clusters were obtained from the Human Protein Atlas (HPA). RNA levels are indicated by normalized transcript per million (nTPM) allowing comparison across all HPA samples. Clusters are assigned based on gene to gene distance and shared nearest neighbor clustering to find clusters of genes with similar expression profiles within the graph. Clusters are then manually annotated using overrepresentation analysis including Gene Ontology, KEGG, Reactome, PanglaoDB, TRRUST and HPA classifications.

Reuse of a pan-cancer dataset (MSV000095036)

MLMarker was trained exclusively on healthy tissue proteomes, making it particularly sensitive to proteomic changes introduced by disease states such as cancer. To assess the extent of this effect, we applied MLMarker to proteomic data from the Tüshaus *et al*. pan-cancer study, which includes 1,220 FFPE tissue samples spanning multiple cancer types and several healthy control samples. We used the pre-processed MassIVE output files (MSV000095036) without re-analysing the raw data. The dataset covered a diverse range of cancers: colorectal cancer (CRC), diffuse large B-cell lymphoma (DLBCL), glioblastoma, HeLa cervical cancer, melanoma, pancreatic ductal adenocarcinoma (PDAC), and oral squamous cell carcinoma (OSCC).

As MLMarker was not trained on cancer-specific proteomes, nor on cell line-derived samples such as HeLa cells, nor on tissues absent from the training set such as skin (melanoma) and organs such as pancreas, we mapped each cancer type to its closest corresponding healthy tissue and applied a penalty factor of 1.

Fig 12a demonstrates the distribution of protein features added to the fixed feature space for the entire dataset with the bulk ranging between 3,500 and 4,500.

Using a strict one-to-one mapping, MLMarker achieved 32.5% accuracy when considering only the top predicted tissue. Expanding to the top five most probable tissues increased the overall accuracy to 50%, with the greatest improvements observed for CRC, healthy lymph node, OSCC, and oral epithelium (OE) (Fig 10). However, oral epithelium lacked a direct match in MLMarker's tissue set and was mapped to the parotid gland, which is functionally distinct from epithelium, potentially explaining the low classification performance. The HeLa cervical cancer cell line was excluded from absolute tissue mapping, as it no longer reflects native tissue but was considered later in comparative analyses. Surprisingly, healthy lymph node performed poorly despite being present in the training data.

Given the heterogeneous nature of cancer proteomes, we expanded our tissue mappings to include related tissues. For example: (i) CRC predictions were considered correct if mapped to colon, small intestine, rectum, duodenum, or appendix; (ii) glioblastoma was mapped to brain and pituitary gland; (iii) OSCC was mapped to parotid gland and tonsil. This broader evaluation strategy significantly improved performance: accuracy increased from 63% (top predicted tissue) to 79% (top five predictions). Notably, samples that previously performed poorly – CRC, healthy lymph node, OSCC, and OE – showed substantial improvements, though CRC remained the weakest performer.

To further validate these findings, we compared MLMarker's predictions against two synthetic control samples: (i) a random sample, generated to mimic the original data distribution; and (ii) a zero-value sample to test the penalty factor performance (Fig 11a).

Both samples clustered distinctly from biological samples, reinforcing the biological relevance of MLMarker's predictions and the efficiency of the penalty factor.

UMAP clustering of the original protein-level expression data revealed clear cancer-type-specific clusters, with some notable exceptions (Fig 11b). Indeed, 30 melanoma samples are metastases from neuronal origin and clustered within the glioblastoma group, reflecting shared protein expression patterns. When restricting the UMAP to only use the MLMarker features, CRC and PDAC became more distinct, but melanoma samples remained interspersed within the glioblastoma cluster. At the prediction level, clustering became less clearly defined, with increased overlap between cancer types, mirroring the original study's top 10 protein UMAP clustering. A striking exception was the HeLa cells dataset, which formed a dense, isolated cluster, indicating a highly distinct proteomic profile, overwhelmingly predicted as B-cells. HeLa cells, originating from cervical cancer tissue, are similar to other tissues in the MLMarker training data, such as endometrium, ovary, oviduct, and placenta, which would presumably resemble cervical tissue more closely than B-cells. To explore this B-cell prediction, we first investigated the impact of missingness on the B-cell prediction by comparing the distribution of added protein features in HeLa samples with that of the entire dataset (Fig. 12a). This comparison revealed a skew towards higher missingness, though HeLa samples were not an outlier. Notably, B-cell similarity did not correlate directly with the number of added features (Fig 12b). We then examined whether the proteins contributing most to the B-cell similarity score were present or absent, given that missing features can also carry information. By separating proteins into present and absent groups and evaluating the direction of their SHAP values, we observed a similar distribution between both sets (Fig 12c). This indicates that the B-cell prediction is not solely driven by absent features. Instead, it suggests that the studied HeLa cell line possessed a proteomic profile more similar to B-cells than to other solid tissues in the reference set, possibly as a result of immortalisation and extensive culturing.

Glioblastoma and melanoma showed no distinct separate clustering. This is notable as MLMarker was never trained on skin, hence melanoma samples cannot be predicted correctly. However, MLMarker did have plentiful training data for brain tissue with some melanoma samples achieving very high brain probabilities.

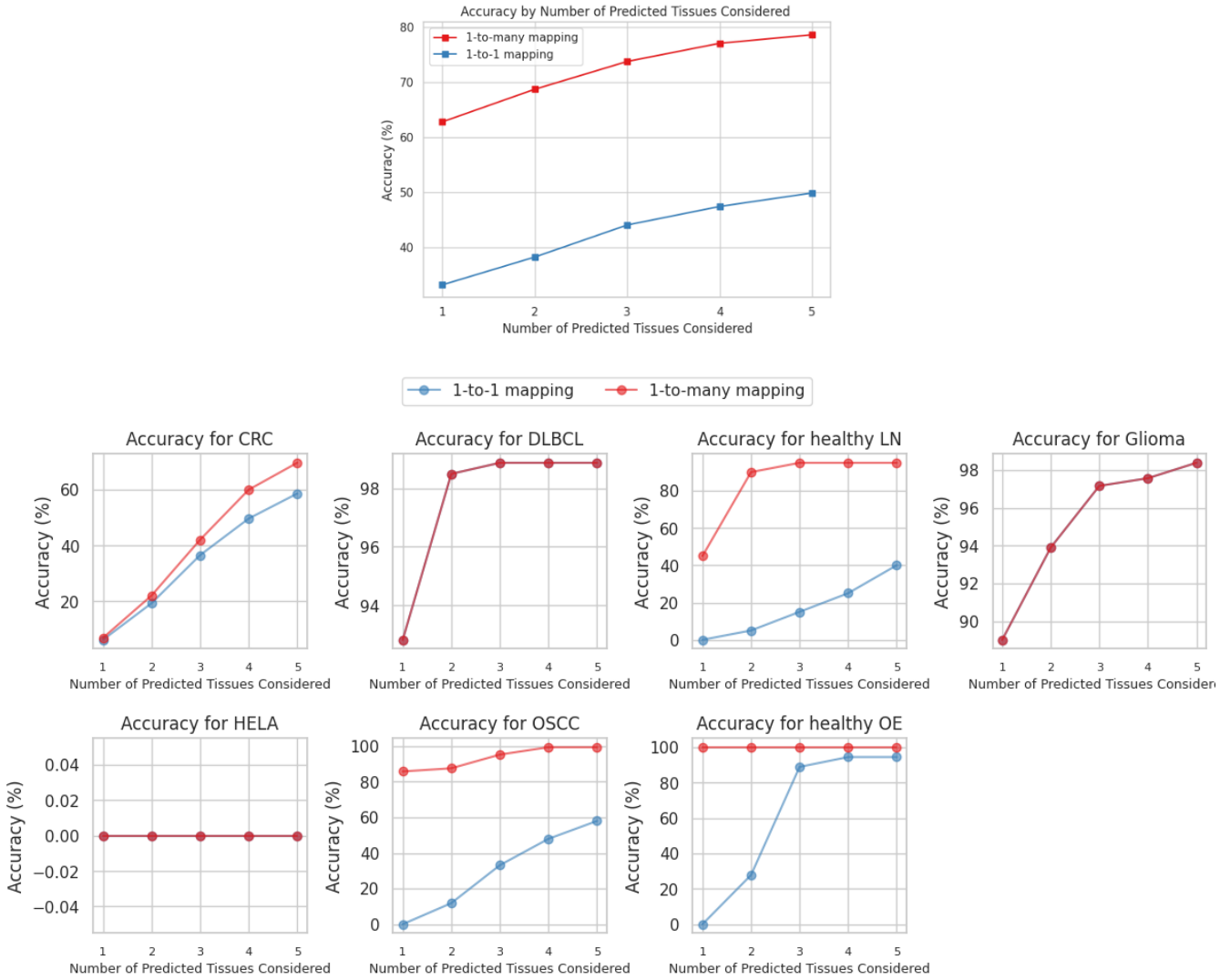


Fig 10 Top: Accuracy of MLMarker predictions on the cancer dataset with a one-to-one mapping (blue) and one-to-many mapping for all cancer types (red) Bottom: accuracy of MLMarker predictions for specific per cancer types in the one-to-one mapping (blue) and one-to-many mapping (red)

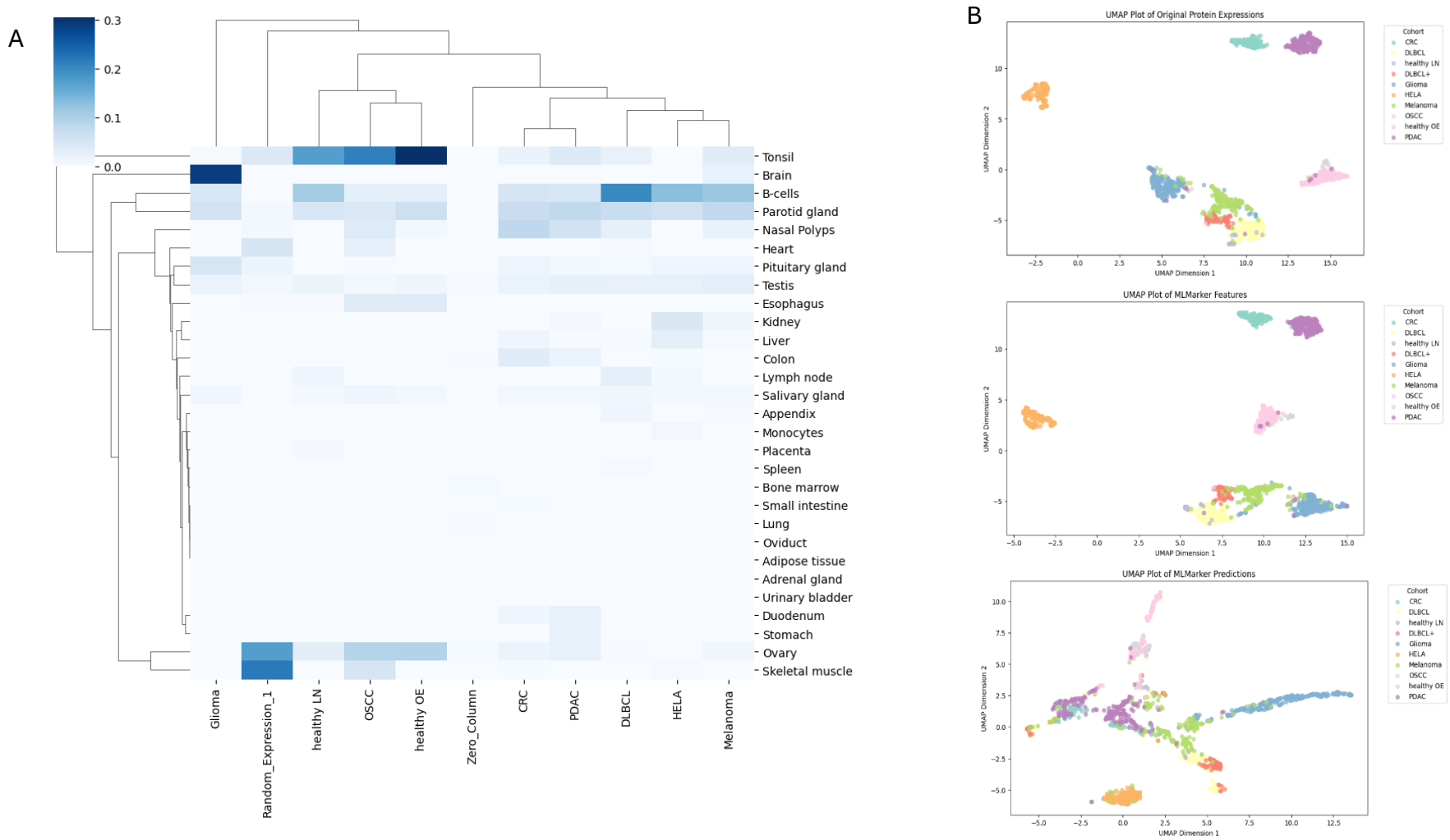


Fig 11A). Dendrogram showing tissue similarity per cancer type, including random and fully zeroed samples (Random_Expression_1 and Zero_column respectively). B) UMAP protein expression of the entire dataset (top panel), protein expression of MLMarker features (middle panel) and tissue similarity values (bottom panel).

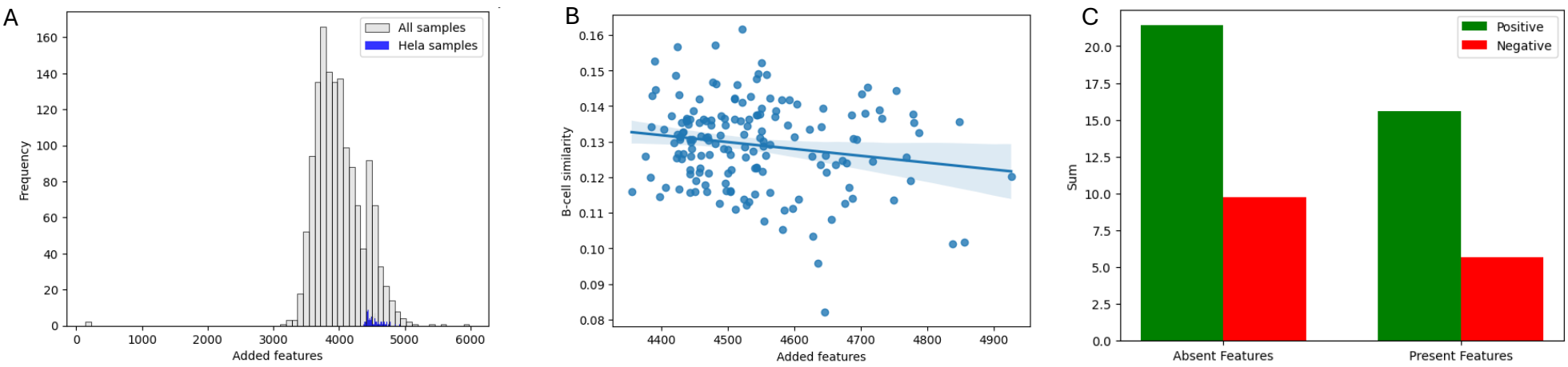


Fig 12 Analysis of B-cell similarity predictions in HeLa samples.

A) Histogram showing the number of protein features added to complete the MLMarker feature space in HeLa samples compared to the full dataset. HeLa samples show a skew toward higher missingness but are not outliers.

B) Scatterplot of B-cell similarity scores versus the number of added protein features in HeLa samples, showing no direct correlation.

C) Distribution of SHAP value directions (positive or negative) for the top B-cell-contributing proteins, separated by presence or absence in the HeLa samples, indicating both present and absent features contribute similarly.

Discussion

We here introduced MLMarker, a data-driven tissue prediction tool that leverages machine learning techniques to identify tissue-specific proteomic signatures, providing a new dimension to the analysis of complex biological data. In essence, MLMarker compares the presented proteome profile to a broad array of publicly available healthy tissue proteomes it was trained upon, providing a boosted DEA. In-depth analysis of the contributing proteins through SHAP analysis clarifies the model's prediction process through Explainable AI and presents a tool for biomarker identification. While originally designed to predict the tissue of origin of samples, MLMarker's ability to uncover underlying proteomic patterns offers a unique approach to generate hypotheses, making it a valuable tool not just for classification, but also for furthering biological understanding in various research contexts.

A prime example of MLMarker's capability is illustrated in the study of cerebral melanoma metastases. The original study stratified tumours based on their response to treatment, but MLMarker presents a proteomic-based classification system, grouping samples based on their similarity to brain tissue. This shift in perspective leads to the hypothesis that certain melanoma tumours, especially those with poor treatment responses, might acquire proteomic characteristics that mimic brain tissue, facilitating their survival and adaptation in the brain's microenvironment. This novel hypothesis was not captured with the original analyses and underscores how MLMarker can reveal new proteomic phenotypes. Following up this observation with differential expression analysis between the clear brain predicted and non-brain predicted sample led to a ten-fold increase in the number of differentially expressed proteins and therefore possibly biologically interesting targets.

By identifying brain-predicted samples, MLMarker revealed that these samples, regardless of their treatment response, retained a distinct set of brain-associated proteins. This suggests that these tumours may be adopting brain-like features through proteomic mimicry, enhancing their capacity to adapt to and survive within the brain environment. In contrast, non-brain-predicted tumours, particularly those with good responses to treatment, were enriched for immune-related pathways, reinforcing the idea that immune activation rather than a specific proteotype may be a key factor in successful treatment responses. These findings provide compelling evidence that tumours exhibiting proteomic similarities to brain tissue might possess an invasive phenotype, enabling them to infiltrate the brain and evade immune surveillance. This hypothesis is further supported by the detection of ten known invasion markers found exclusively in brain-predicted samples and absent in non-brain-predicted tumours. This distinction emphasizes the connection between proteomic profiles and the invasive potential of tumours.

In addition to its role in identifying proteomic features associated with tumour invasiveness, MLMarker reinforces the concept that cancer proteomes strongly diverge from their healthy counterpart[27,29] and some deviate more than others. An unexpected finding in the pan-cancer study was the clustering of melanoma samples with glioblastoma, indicating that certain melanoma cases displayed proteomic patterns similar to those of glioblastoma. Although some of these were known neuronal metastases, this was not the case for each melanoma and requires further investigation into the specific proteins responsible for this observation.

One of the most valuable features of MLMarker is its ability to handle sparse or noisy data, an issue that often arises in proteomics studies. By incorporating a penalty factor and adjusting tissue mappings, MLMarker was able to improve classification accuracy and sensitivity, even when faced with missing or incomplete protein data. This capability is particularly relevant for clinical applications, where liquid biopsies and other biofluid samples are increasingly being used for cancer detection and monitoring. These samples, by nature, contain a higher degree of missing data compared to traditional tissue biopsies, making MLMarker's robustness in handling incomplete data a significant advantage. In fact, MLMarker was able to identify the tissue of origin for tissue-leakage proteins (TLPs) in biofluids, offering a new approach for detecting and monitoring cancer through biofluid analysis. Interestingly, MLMarker also detects a notable presence of liver-associated proteins, a phenomenon previously observed in both brain tissue and CSF[38]. However, rather than misclassifying these samples, MLMarker correctly recognizes that the co-occurrence of brain proteins with these liver proteins still indicates a neuronal origin rather than hepatic contamination. This further underscores MLMarker's ability to handle complex tissue compositions and its adaptability to sample types beyond those in its training data, provided that appropriate adjustments, such as penalty factors, are applied.

HeLa cells clustered separately from both healthy and cancerous tissues, reflecting a distinct proteomic signature. Despite their cervical origin, these samples were not predicted as any possibly related tissues. Instead, they were consistently predicted as B-cells, raising the question of whether the HeLa proteome has diverged too far from its original tissue context, resulting in assignment to the nearest isolated cell class. This opens future opportunities to study cell line proteome profiles from a tissue point of view.

By leveraging a fine-grained Random Forest model, MLMarker does not force overly confident classifications but instead allows for some degree of ambiguity in its predictions. As a result, the expected tissue may not always rank as the top prediction, reflecting the inherent complexity and overlap in proteomic profiles across tissues. Rather than being a limitation, this characteristic provides a biologically meaningful perspective, where examining the top n predicted tissues can reveal underlying

relationships between sample proteomes and multiple tissue types, offering deeper insights into tissue similarity and potential cross-tissue influences.

Looking to the future, the development of MLMarker continues to expand its potential applications. Currently available as a Python package and a Streamlit app, MLMarker is designed to bridge the gap between wet-lab and dry-lab scientists. This makes it a versatile tool that can be used by a wide range of researchers, from experimental biologists to computational scientists, fostering collaboration across disciplines. However, as with any tool, users must be aware of its limitations. To aid with this, comprehensive documentation has been provided, which offers further context and guidance on using the tool effectively.

In the realm of clinical research, MLMarker holds significant promise for identifying tissue origins of metastases. The tool's capacity to classify tissues based on their proteomic profiles rather than conventional treatment-response criteria opens up new avenues for understanding metastatic behaviour. Additionally, MLMarker could play a pivotal role in overcoming challenges in single-cell proteomics, where hypothesis generation is often limited. By replacing traditional clustering-based differential expression analysis (DEA) with MLMarker's orthogonal classification method, the tool could avoid the circularity issue known as double dipping, in which the same clusters are used for both generating hypotheses and performing subsequent analyses. This could streamline the workflow and lead to more accurate, unbiased results in proteomics studies.

Ongoing improvements to MLMarker aim to address challenges related to missing values and further expand the dataset used in its training. However, integrating more public data remains challenging due to the continuing lack of comprehensive metadata in publicly available proteomics datasets. The importance of metadata cannot be overstated, as it is essential for ensuring reproducibility and enabling the reuse of datasets across research projects. The development of metadata standards, such as the SDRF-Proteomics metadata standard and tools like lesSDRF[39] are critical steps forward in addressing this issue. Widespread adoption of these standards will not only improve data quality and accessibility but also accelerate the development of data-driven methods like MLMarker.

In conclusion, MLMarker represents a significant innovation in the field of proteomics, providing a data-driven tool for hypothesis generation and improving the interpretation of complex proteomic data. Its abilities to handle missing data, to classify tissues based on proteomic profiles, and to generate new biological insights make it a versatile and indispensable tool in both basic research and clinical applications. As MLMarker continues to evolve, it will undoubtedly contribute to a deeper understanding of cancer biology and other complex diseases, paving the way for more precise and personalized therapeutic strategies. The continued development and adoption of standards for proteomics data will further enhance the tool's impact, ensuring that it remains a key

resource for researchers seeking to unlock the full potential of proteomics in biomedical science.

Conclusion

MLMarker moves beyond conventional classification and differential analysis by introducing a continuous, probabilistic framework for assessing tissue similarity in proteomics data. Instead of relying on predefined groups, the tool compares each sample against a reference atlas of healthy tissue proteomes, producing interpretable similarity scores that reflect underlying biology. The model operates within a fixed protein feature space and was trained on normalized expression data from 34 healthy tissues. A penalty factor corrects for the impact of missing proteins, making MLMarker robust to sparsity and suitable for complex samples such as biofluids or tumour tissues with partial proteome coverage. SHAP values provide a transparent link between input features and predicted tissue similarities, enabling the identification of proteins that drive classification.

MLMarker was particularly effective in creating new biological hypotheses. In cerebral melanoma metastases, MLMarker identified a subset of tumours with high similarity to brain tissue, pointing to proteomic adaptation to the brain microenvironment, an insight not captured by conventional analysis. The application on biofluid and cancer data, very distinct from the original healthy tissue training data demonstrated the model's robustness and innovation. Gene set enrichment based on SHAP-ranked proteins further validated the biological relevance of model outputs.

As proteomics advances toward higher resolution and broader application, the ability to quantify tissue similarity in an interpretable and data-resilient manner becomes increasingly important. MLMarker addresses this need, offering a scalable tool for hypothesis generation and biological interpretation across diverse experimental contexts.

Materials and methods

Public datasets used for MLMarker application

Cerebral Melanoma Metastases Proteomics (PXD007592)

The dataset PXD007592 from Zila *et al*., initially generated to investigate proteomic differences between good and poor responders to MAP kinase inhibitors (MAPKi) in cerebral melanoma metastases, was used in this study. Raw mass spectrometry data files were downloaded from the PRIDE repository and reprocessed using an updated computational workflow (outlined below in the Reprocessing Pipeline section). Sample metadata, including the treatment response classification (good vs. poor responders), was inferred directly from the file names provided in the original dataset.

Cerebrospinal Fluid (CSF) Proteome Dataset (PXD008029)

To assess the applicability of MLMarker to biofluid proteomics, cerebrospinal fluid (CSF) samples from the Yamana dataset PXD008029 were selected for analysis. Sample metadata was extracted from the file names provided in the original dataset.

Pan-Cancer Formalin-Fixed Paraffin-Embedded (FFPE) Proteomics Dataset (MSV000095036)

The dataset from Tüshaus *et al*. consists of proteomic profiles for 1,220 FFPE tumor samples from multiple cancer types. Unlike the other datasets, raw data were not reprocessed for this study. Instead, processed output files from the original study, which are publicly available on MassIVE, were used directly for MLMarker analysis. Sample metadata, including cancer type and tumor classification, was retrieved from the mapping file provided by the original study authors in the MassIVE files.

Reprocessing pipeline

For datasets requiring reprocessing, specifically PXD007592 (melanoma) and PXD008029 (CSF), a standardized computational workflow was applied to ensure consistency across datasets.

Protein Identification: Ionbot[33] (v0.11.0) was employed for open modification searches. The Uniprot human reference proteome (reviewed) was used for sequence matching. Carbamidomethylation of cysteine (C+57.021464 Da) was set as a fixed modification, while methionine oxidation (M+15.994915 Da) was set as a variable modification. The precursor mass tolerance was ±10 ppm, and the fragment mass tolerance was ±20 ppm. Peptide-spectrum match (PSM) false discovery rate (FDR) was controlled to be < 1%.

Protein Quantification: FlashLFQ [34](v1.0.3) was used for label-free quantification. The "match-between-runs" option was disabled to ensure independent quantification per sample.

Differential Expression Analysis: MSqRob2[35](v1.4.0) was used for robust statistical analysis of differential expression. Starting from quantified peptides, robustly summarized proteins detected in at least four samples were retained for analysis. A hurdle model was employed for statistical modeling. Pairwise contrasts were defined for: Good vs. poor responders (melanoma dataset) and Brain-predicted vs. non-brain-predicted samples (MLMarker-derived groups).

Interpretation analyses

MLMarker[26] (v0.1.4), a Random Forest-based tissue classifier trained on healthy tissue proteomes, was applied to all datasets in this study. The input consisted of normalized protein spectral abundances, which were used for tissue similarity prediction. Tissue probability scores were then assigned to each sample, indicating the most probable tissue of origin. We used the quantitative model with 5,979 protein features.

Fixed feature space

MLMarker relies on a fixed feature space, meaning it can only utilize proteins present in the training dataset. Missing proteins are assigned a value of zero, which may skew predictions, especially when sample types contain proteins not represented in the training set. In these cases, the model predicts using available proteins, but missing data can lead to biased or inconsistent tissue predictions. MLMarker relies on a fixed feature space, meaning it can only utilize proteins present in the training dataset. Missing proteins are assigned a value of zero, which may skew predictions, especially when sample types contain proteins not represented in the training set. In these cases, the model predicts using available proteins, but missing data can lead to biased or inconsistent tissue predictions.

Penalty factor

To mitigate the effect of missing proteins, MLMarker includes a penalty factor. This factor adjusts SHAP values for absent proteins, which could otherwise distort tissue classification. Missing proteins that contribute non-zero SHAP values are penalized using a baseline zeroed-out sample. The penalty factor is applied to these proteins, reducing their influence on the prediction. Present proteins remain unaffected and absent proteins with no contribution are left unadjusted. This correction improves the reliability of SHAP values and tissue classification in samples with missing data. The adjusted SHAP values are computed by separating present and absent proteins and penalizing contributing absent proteins while preserving the contributions of detected proteins.

Uniform Manifold Approximation and Projection (UMAP) was applied to visualize the data using the umap package v0.1.1.

SHAP (SHapley Additive exPlanations) values were calculated using the shap package (v0.46.0) to interpret the contributions of individual proteins to tissue classification predictions. Positive SHAP values were associated with proteins that increased the likelihood of a given tissue classification, while negative SHAP values indicated proteins that decreased the likelihood of the classification.

Gene Ontology (GO) enrichment analysis was performed using the GProfiler tool (v1.0.0) to identify enriched biological processes, molecular functions, and cellular components. KEGG and Reactome pathways, as well as tissue-specific expression data from the Human Protein Atlas, were also incorporated into the analysis. Proteins ranked by SHAP feature importance were used as input for this analysis. Benjamini-Hochberg correction was used to obtain the adjusted p-value.

Correlation analysis between adjusted SHAP values and adjusted p-values was performed using Pearson correlation coefficient.

Data and code Availability

All datasets used in this study are publicly available via ProteomeXchange (PXD007592, PXD008029) and MassIVE (MSV000095036). MLMarker predictions, differential expression results, and SHAP-derived protein lists are available in the Supplementary Data & Github repository for the manuscript analyses, pip package and streamlit application respectively github.com/TineClaeys/MLMarker-manuscript; github.com/TineClaeys/MLMarker; github.com/TineClaeys/MLMarker-streamlit. The streamlit application is available through mlmarker.streamlit.app.

References

(1) Krieger, J. R.; Wybenga-Groot, L. E.; Tong, J.; Bache, N.; Tsao, M. S.; Moran, M. F. Evosep One Enables Robust Deep Proteome Coverage Using Tandem Mass Tags While Significantly Reducing Instrument Time. *J. Proteome Res.* **2019**, *18* (5), 2346–2353. https://doi.org/10.1021/acs.jproteome.9b00082.

(2) Ye, Z.; Sabatier, P.; van der Hoeven, L.; Lechner, M. Y.; Phlairaharn, T.; Guzman, U. H.; Liu, Z.; Huang, H.; Huang, M.; Li, X.; Hartlmayr, D.; Izaguirre, F.; Seth, A.; Joshi, H. J.; Rodin, S.; Grinnemo, K.-H.; Hørning, O. B.; Bekker-Jensen, D. B.; Bache, N.; Olsen, J. V. Enhanced Sensitivity and Scalability with a Chip-Tip Workflow Enables Deep Single-Cell Proteomics. *Nat. Methods* **2025**, 1–11. https://doi.org/10.1038/s41592-024-02558-2.

(3) Wang, F.; Veth, T.; Kuipers, M.; Altelaar, M.; Stecker, K. E. Optimized Suspension Trapping Method for Phosphoproteomics Sample Preparation. *Anal. Chem.* **2023**, *95* (25), 9471–9479. https://doi.org/10.1021/acs.analchem.3c00324.

(4) Ctortecka, C.; Clark, N. M.; Boyle, B. W.; Seth, A.; Mani, D. R.; Udeshi, N. D.; Carr, S. A. Automated Single-Cell Proteomics Providing Sufficient Proteome Depth to Study Complex Biology beyond Cell Type Classifications. *Nat. Commun.* **2024**, *15* (1), 5707. https://doi.org/10.1038/s41467-024-49651-w.

(5) Bubis, J. A.; Arrey, T. N.; Damoc, E.; Delanghe, B.; Slovakova, J.; Sommer, T. M.; Kagawa, H.; Pichler, P.; Rivron, N.; Mechtler, K.; Matzinger, M. Challenging the Astral Mass Analyzer to Quantify up to 5,300 Proteins per Single Cell at Unseen Accuracy to Uncover Cellular Heterogeneity. *Nat. Methods* **2025**, 1–10. https://doi.org/10.1038/s41592-024-02559-1.

(6) Stewart, H. I.; Grinfeld, D.; Giannakopulos, A.; Petzoldt, J.; Shanley, T.; Garland, M.; Denisov, E.; Peterson, A. C.; Damoc, E.; Zeller, M.; Arrey, T. N.; Pashkova, A.; Renuse, S.; Hakimi, A.; Kühn, A.; Biel, M.; Kreutzmann, A.; Hagedorn, B.; Colonius, I.; Schütz, A.; Stefes, A.; Dwivedi, A.; Mourad, D.; Hoek, M.; Reitemeier, B.; Cochems, P.; Kholomeev, A.; Ostermann, R.; Quiring, G.; Ochmann, M.; Möhring, S.; Wagner, A.; Petker, A.; Kanngiesser, S.; Wiedemeyer, M.; Balschun, W.; Hermanson, D.; Zabrouskov, V.; Makarov, A. A.; Hock, C. Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis. *Anal. Chem.* **2023**, *95* (42), 15656–15664. https://doi.org/10.1021/acs.analchem.3c02856.

(7) Declercq, A.; Bouwmeester, R.; Hirschler, A.; Carapito, C.; Degroeve, S.; Martens, L.; Gabriels, R. MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell. Proteomics MCP* **2022**, *21* (8), 100266. https://doi.org/10.1016/j.mcpro.2022.100266.

(8) Demichev, V.; Szyrwiel, L.; Yu, F.; Teo, G. C.; Rosenberger, G.; Niewienda, A.; Ludwig, D.; Decker, J.; Kaspar-Schoenefeld, S.; Lilley, K. S.; Mülleder, M.; Nesvizhskii, A. I.; Ralser, M. Dia-PASEF Data Analysis Using FragPipe and DIA-NN for Deep Proteomics of Low Sample Amounts. *Nat. Commun.* **2022**, *13* (1), 3944. https://doi.org/10.1038/s41467-022-31492-0.

(9) Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J. Proteome Res.* **2023**, *22* (11), 3652–3659. https://doi.org/10.1021/acs.jproteome.3c00486.

(10) Teo, G. C.; Polasky, D. A.; Yu, F.; Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J. Proteome Res.* **2021**, *20* (1), 498–505. https://doi.org/10.1021/acs.jproteome.0c00544.
(11) Domon, B.; Aebersold, R. Challenges and Opportunities in Proteomics Data Analysis*. *Mol. Cell. Proteomics* **2006**, *5* (10), 1921–1926. https://doi.org/10.1074/mcp.R600012-MCP200.
(12) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214. https://doi.org/10.1038/nchembio.2576.
(13) Peng, H.; Wang, H.; Kong, W.; Li, J.; Goh, W. W. B. Optimizing Differential Expression Analysis for Proteomics Data via High-Performing Rules and Ensemble Inference. *Nat. Commun.* **2024**, *15* (1), 3922. https://doi.org/10.1038/s41467-024-47899-w.
(14) Langley, S. R.; Mayr, M. Comparative Analysis of Statistical Methods Used for Detecting Differential Expression in Label-Free Mass Spectrometry Proteomics. *J. Proteomics* **2015**, *129*, 83–92. https://doi.org/10.1016/j.jprot.2015.07.012.
(15) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.
(16) Perez-Riverol, Y.; Bandla, C.; Kundu, D. J.; Kamatchinathan, S.; Bai, J.; Hewapathirana, S.; John, N. S.; Prakash, A.; Walzer, M.; Wang, S.; Vizcaíno, J. A. The PRIDE Database at 20 Years: 2025 Update. *Nucleic Acids Res.* **2025**, *53* (D1), D543–D553. https://doi.org/10.1093/nar/gkae1011.
(17) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas Project. *Nucleic Acids Res.* **2006**, *34* (suppl_1), D655–D658. https://doi.org/10.1093/nar/gkj040.
(18) Schmidt, T.; Samaras, P.; Frejno, M.; Gessulat, S.; Barnert, M.; Kienegger, H.; Krcmar, H.; Schlegl, J.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. ProteomicsDB. *Nucleic Acids Res.* **2018**, *46* (D1), D1271–D1281. https://doi.org/10.1093/nar/gkx1029.
(19) Ramasamy, P.; Turan, D.; Tichshenko, N.; Hulstaert, N.; Vandermarliere, E.; Vranken, W.; Martens, L. Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context. *J. Proteome Res.* **2020**, *19* (8), 3478–3486. https://doi.org/10.1021/acs.jproteome.0c00306.
(20) Degroeve, S.; Martens, L. MS2PIP: A Tool for MS/MS Peak Intensity Prediction. *Bioinforma. Oxf. Engl.* **2013**, *29* (24), 3199–3203. https://doi.org/10.1093/bioinformatics/btt544.

(21) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. *Nat. Methods* **2021**, *18* (11), 1363–1369. https://doi.org/10.1038/s41592-021-01301-5.
(22) Declercq, A.; Devreese, R.; Scheid, J.; Jachmann, C.; Van Den Bossche, T.; Preikschat, A.; Gomez-Zepeda, D.; Rijal, J. B.; Hirschler, A.; Krieger, J. R.; Srikumar, T.; Rosenberger, G.; Martelli, C.; Trede, D.; Carapito, C.; Tenzer, S.; Walz, J. S.; Degroeve, S.; Bouwmeester, R.; Martens, L.; Gabriels, R. TIMS2Rescore: A Data Dependent Acquisition-Parallel Accumulation and Serial Fragmentation-Optimized Data-Driven Rescoring Pipeline Based on MS2Rescore. *J. Proteome Res.* **2025**, *24* (3), 1067–1076. https://doi.org/10.1021/acs.jproteome.4c00609.
(23) Robles, J.; Prakash, A.; Vizcaíno, J. A.; Casal, J. I. Integrated Meta-Analysis of Colorectal Cancer Public Proteomic Datasets for Biomarker Discovery and Validation. *PLOS Comput. Biol.* **2024**, *20* (1), e1011828-.
(24) Camacho, O. J. M.; Ramsbottom, K. A.; Prakash, A.; Sun, Z.; Riverol, Y. P.; Bowler-Barnett, E.; Martin, M.; Fan, J.; Deutsch, E. W.; Vizcaíno, J. A.; Jones, A. R. Phosphorylation in the <Em>Plasmodium Falciparum</Em> Proteome: A Meta-Analysis of Publicly Available Data Sets. *bioRxiv* **2023**, 2023.11.20.567785. https://doi.org/10.1101/2023.11.20.567785.
(25) Ochoa, D.; Jarnuczak, A. F.; Viéitez, C.; Gehre, M.; Soucheray, M.; Mateus, A.; Kleefeldt, A. A.; Hill, A.; Garcia-Alonso, L.; Stein, F.; Krogan, N. J.; Savitski, M. M.; Swaney, D. L.; Vizcaíno, J. A.; Noh, K.-M.; Beltrao, P. The Functional Landscape of the Human Phosphoproteome. *Nat. Biotechnol.* **2020**, *38* (3), 365–373. https://doi.org/10.1038/s41587-019-0344-3.
(26) Claeys, T.; Menu, M.; Bouwmeester, R.; Gevaert, K.; Martens, L. Machine Learning on Large-Scale Proteomics Data Identifies Tissue and Cell-Type Specific Proteins. *J. Proteome Res.* **2023**. https://doi.org/10.1021/acs.jproteome.2c00644.
(27) Kosti, I.; Jain, N.; Aran, D.; Butte, A. J.; Sirota, M. Cross-Tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.* **2016**, *6*, 24799. https://doi.org/10.1038/srep24799.
(28) De Wilde, J.; Van Paemel, R.; De Koker, A.; Roelandt, S.; Van de Velde, S.; Callewaert, N.; Van Dorpe, J.; Creytens, D.; De Wilde, B.; De Preter, K. A Fast, Affordable, and Minimally Invasive Diagnostic Test for Cancer of Unknown Primary Using DNA Methylation Profiling. *Lab. Invest.* **2024**, *104* (8), 102091. https://doi.org/10.1016/j.labinv.2024.102091.
(29) Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; Aguet, F.; Anand, S.; Ardlie, K. G.; Gabriel, S.; Getz, G.; Graubert, A.; Hadley, K.; Handsaker, R. E.; Huang, K. H.; Kashin, S.; MacArthur, D. G.; Meier, S. R.; Nedzel, J. L.; Nguyen, D. Y.; Segrè, A. V.; Todres, E.; Balliu, B.; Barbeira, A. N.; Battle, A.; Bonazzola, R.; Brown, A.; Brown, C. D.; Castel, S. E.; Conrad, D.; Cotter, D. J.; Cox, N.; Das, S.; de Goede, O. M.; Dermitzakis, E. T.; Engelhardt, B. E.; Eskin, E.; Eulalio, T. Y.; Ferraro, N. M.; Flynn, E.; Fresard, L.; Gamazon, E. R.; Garrido-Martín, D.; Gay, N. R.; Guigó, R.; Hamel, A. R.; He, Y.; Hoffman, P. J.; Hormozdiari, F.; Hou, L.; Im, H. K.; Jo, B.; Kasela, S.; Kellis, M.; Kim-Hellmuth, S.; Kwong, A.; Lappalainen, T.; Li, X.; Liang, Y.; Mangul, S.; Mohammadi, P.; Montgomery, S. B.; Muñoz-Aguirre, M.; Nachun, D. C.; Nobel, A. B.; Oliva, M.; Park, Y. S.; Park, Y.; Parsana, P.; Reverter, F.; Rouhana, J. M.; Sabatti, C.; Saha, A.; Skol, A. D.; Stephens, M.; Stranger, B. E.; Strober, B. J.; Teran, N. A.; Viñuela, A.; Wang, G.; Wen, X.; Wright, F.; Wucher, V.;

Zou, Y.; Ferreira, P. G.; Li, G.; Melé, M.; Yeger-Lotem, E.; Barcus, M. E.; Bradbury, D.; Krubit, T.; McLean, J. A.; Qi, L.; Robinson, K.; Roche, N. V.; Smith, A. M.; Sobin, L.; Tabor, D. E.; Undale, A.; Bridge, J.; Brigham, L. E.; Foster, B. A.; Gillard, B. M.; Hasz, R.; Hunter, M.; Johns, C.; Johnson, M.; Karasik, E.; Kopen, G.; Leinweber, W. F.; McDonald, A.; Moser, M. T.; Myer, K.; Ramsey, K. D.; Roe, B.; Shad, S.; Thomas, J. A.; Walters, G.; Washington, M.; Wheeler, J.; Jewell, S. D.; Rohrer, D. C.; Valley, D. R.; Davis, D. A.; Mash, D. C.; Branton, P. A.; Barker, L. K.; Gardiner, H. M.; Mosavel, M.; Siminoff, L. A.; Flicek, P.; Haeussler, M.; Juettemann, T.; Kent, W. J.; Lee, C. M.; Powell, C. C.; Rosenbloom, K. R.; Ruffier, M.; Sheppard, D.; Taylor, K.; Trevanion, S. J.; Zerbino, D. R.; Abell, N. S.; Akey, J.; Chen, L.; Demanelis, K.; Doherty, J. A.; Feinberg, A. P.; Hansen, K. D.; Hickey, P. F.; Jasmine, F.; Kaul, R.; Kibriya, M. G.; Li, J. B.; Li, Q.; Linder, S. E.; Pierce, B. L.; Rizzardi, L. F.; Smith, K. S.; Stamatoyannopoulos, J.; Tang, H.; Carithers, L. J.; Guan, P.; Koester, S. E.; Little, A. R.; Moore, H. M.; Nierras, C. R.; Rao, A. K.; Vaught, J. B.; Volpi, S.; Snyder, M. P. A Quantitative Proteome Map of the Human Body. *Cell* **2020**, *183* (1), 269-283.e19. https://doi.org/10.1016/j.cell.2020.08.036.

(30) Zila, N.; Bileck, A.; Muqaku, B.; Janker, L.; Eichhoff, O. M.; Cheng, P. F.; Dummer, R.; Levesque, M. P.; Gerner, C.; Paulitschke, V. Proteomics-Based Insights into Mitogen-Activated Protein Kinase Inhibitor Resistance of Cerebral Melanoma Metastases. *Clin. Proteomics* **2018**, *15* (1), 13. https://doi.org/10.1186/s12014-018-9189-x.

(31) Rapid and Deep Profiling of Human Induced Pluripotent Stem Cell Proteome by One-Shot NanoLC–MS/MS Analysis with Meter-Scale Monolithic Silica Columns | Journal of Proteome Research.

(32) Tüshaus, J.; Eckert, S.; Schliemann, M.; Zhou, Y.; Pfeiffer, P.; Halves, C.; Fusco, F.; Weigel, J.; Hönikl, L.; Butenschön, V.; Todorova, R.; Rauert-Wunderlich, H.; The, M.; Rosenwald, A.; Heinemann, V.; Holch, J.; Steiger, K.; Delbridge, C.; Meyer, B.; Weichert, W.; Mogler, C.; Kuhn, P.-H.; Kuster, B. Towards Routine Proteome Profiling of FFPE Tissue: Insights from a 1,220-Case Pan-Cancer Study. *EMBO J.* **2025**, *44* (1), 304–329. https://doi.org/10.1038/s44318-024-00289-w.

(33) Degroeve, S.; Gabriels, R.; Velghe, K.; Bouwmeester, R.; Tichshenko, N.; Martens, L. Ionbot: A Novel, Innovative and Sensitive Machine Learning Approach to LC-MS/MS Peptide Identification. *bioRxiv* **2022**, 2021.07.02.450686. https://doi.org/10.1101/2021.07.02.450686.

(34) Millikin, R. J.; Solntsev, S. K.; Shortreed, M. R.; Smith, L. M. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17* (1), 386–391. https://doi.org/10.1021/acs.jproteome.7b00608.

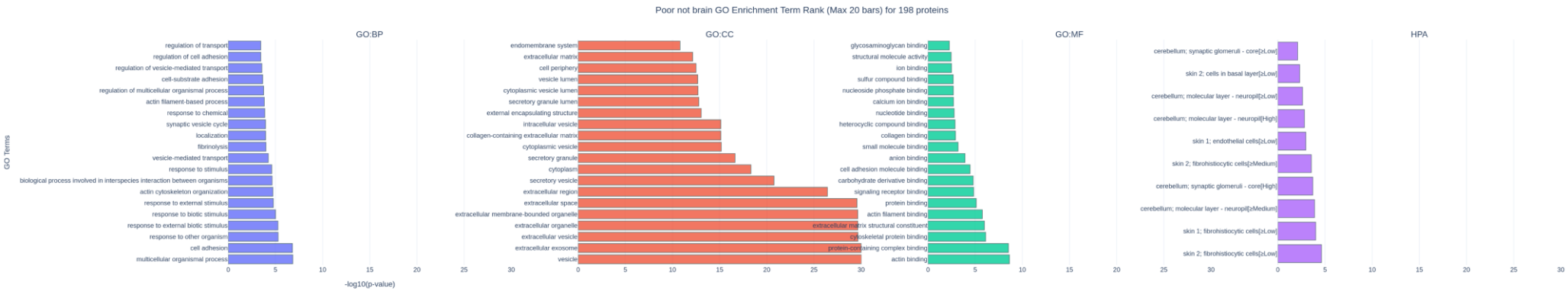
(35) Goeminne, L. J. E.; Sticker, A.; Martens, L.; Gevaert, K.; Clement, L. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Anal. Chem.* **2020**, *92* (9), 6278–6287. https://doi.org/10.1021/acs.analchem.9b04375.

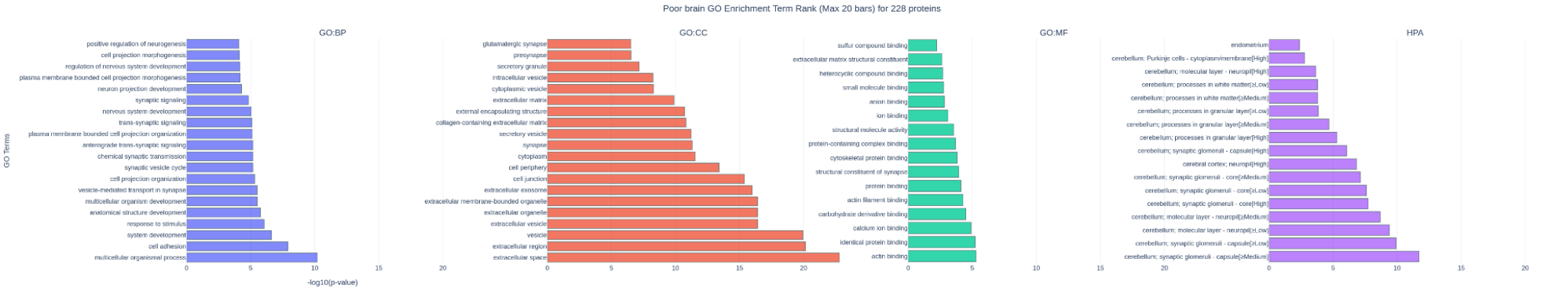
(36) Pouliquen, D. L.; Boissard, A.; Coqueret, O.; Guette, C. Biomarkers of Tumor Invasiveness in Proteomics (Review). *Int. J. Oncol.* **2020**, *57* (2), 409–432. https://doi.org/10.3892/ijo.2020.5075.

(37) Macron, C.; Lane, L.; Núñez Galindo, A.; Dayon, L. Identification of Missing Proteins in Normal Human Cerebrospinal Fluid. *J. Proteome Res.* **2018**, *17* (12), 4315–4319. https://doi.org/10.1021/acs.jproteome.8b00194.

(38) Fang, Q.; Strand, A.; Law, W.; Faca, V. M.; Fitzgibbon, M. P.; Hamel, N.; Houle, B.; Liu, X.; May, D. H.; Poschmann, G.; Roy, L.; Stühler, K.; Ying, W.; Zhang, J.; Zheng, Z.; Bergeron, J. J. M.; Hanash, S.; He, F.; Leavitt, B. R.; Meyer, H. E.; Qian, X.; McIntosh, M. W. Brain-Specific Proteins Decline in the Cerebrospinal Fluid of Humans with Huntington Disease. *Mol. Cell. Proteomics MCP* **2009**, *8* (3), 451–466. https://doi.org/10.1074/mcp.M800231-MCP200.
(39) Claeys, T.; Van Den Bossche, T.; Perez-Riverol, Y.; Gevaert, K.; Vizcaíno, J. A.; Martens, L. lesSDRF Is More: Maximizing The Value Of Proteomics Data Through Streamlined Metadata Annotation. *Nat. Commun.* **2023**, *14* (6473). https://dx.doi.org/10.21203/rs.3.rs-2937726/v1.

Supplementary materials





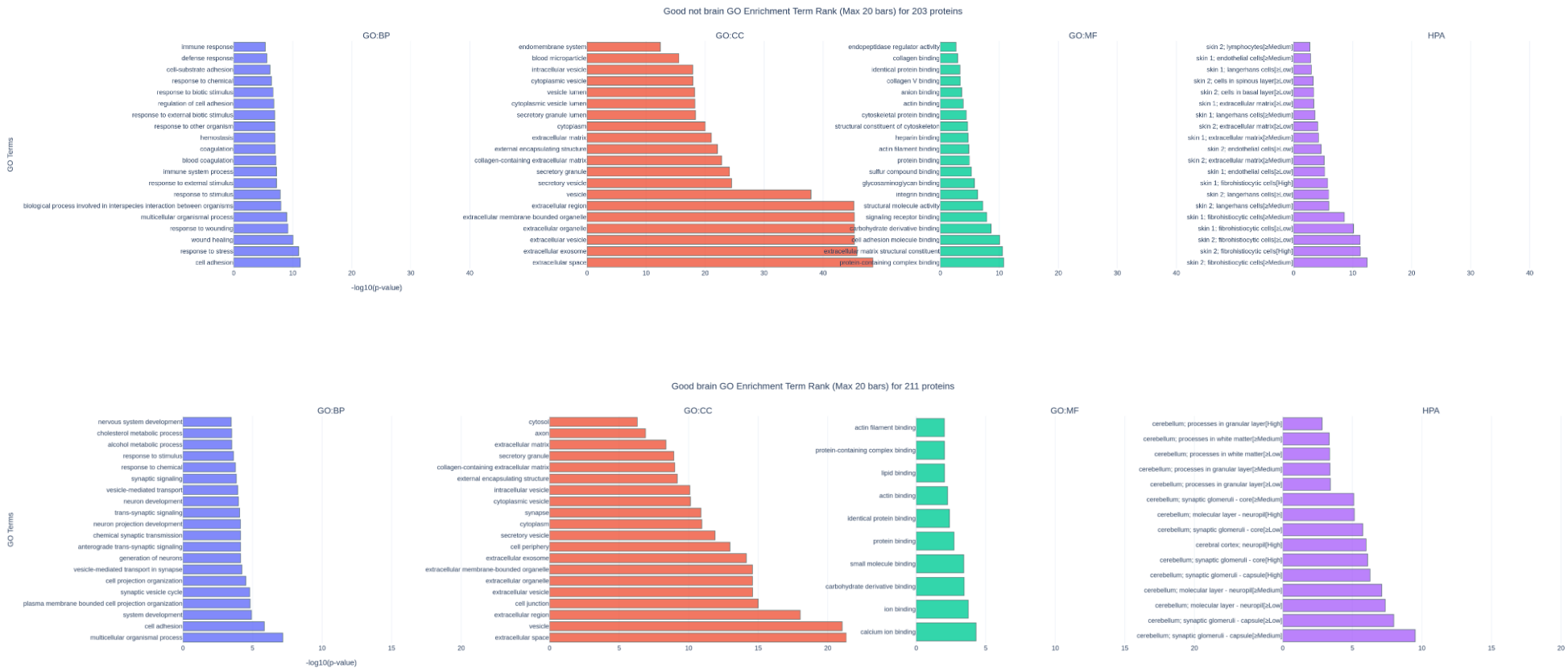


Fig S1: Gene ontology enrichment of the top 1% highest ranking proteins for the four subpopulations at p-value of 1%.

Fig S2: Gene ontology enrichment of the 241 differentially expressed proteins upregulated in the brain predicted samples (top) and not-brain predicted samples (bottom) at a p-value of 1%.

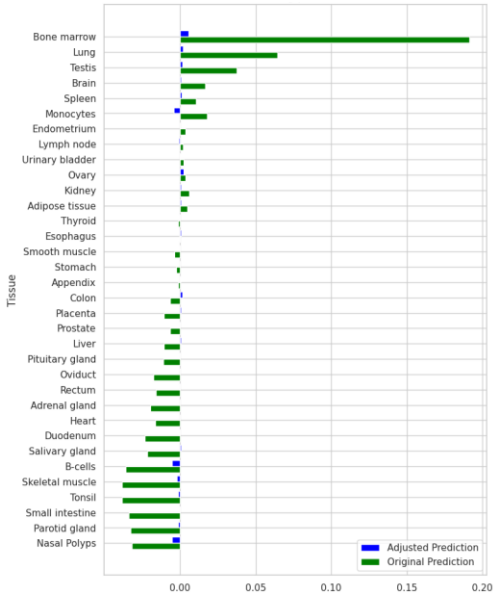


Fig S3: Bar plot showing tissue similarity scores for a fully zeroed sample. Original predictions are shown in green; adjusted predictions using the penalty factor are shown in blue. The application of the penalty factor results in a strong reduction in prediction confidence across all tissues.

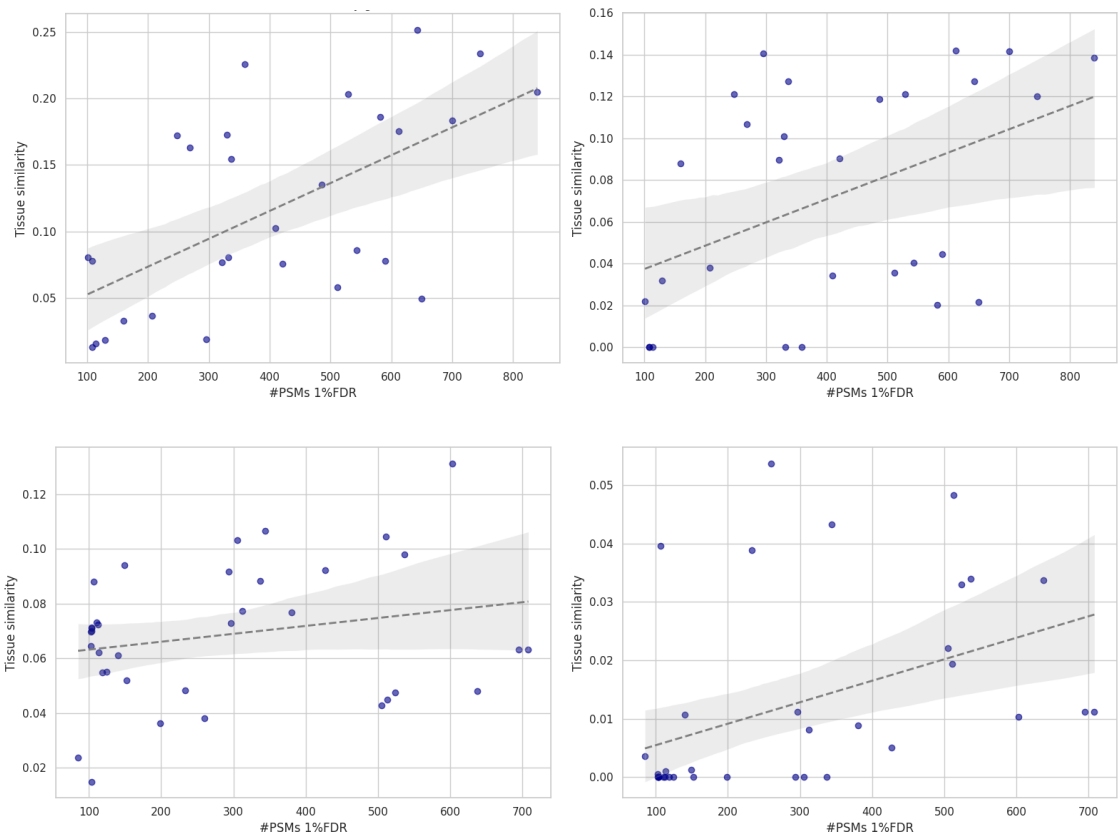


Fig S4: Scatterplot showing the relationship between tissue similarity and PSM counts for pituitary gland (top left), brain in the CSF samples (top right) and for adipose tissue (bottom left) and monocytes in plasma (bottom right). The x-axis represents the number

of PSMs passing the 1% FDR threshold. A linear regression trendline indicates that an increase in PSMs often leads to greater certainty in tissue similarity predictions.

Table S1