“In God we trust. All others must bring data.”

William Edwards Deming

University of Szeged
Albert Szent-Györgyi Medical School
Doctoral School of Interdisciplinary Medicine

SMALL VESICLES, GREAT VALUE: MACHINE LEARNING ANALYSIS OF MOLECULAR FINGERPRINT IN EXTRACELLULAR VESICLES FOR TUMOR DIAGNOSTIC PURPOSES

Ph.D. Thesis

Mátyás Bukva

Supervisor

Krisztina Buzás, Ph.D.

Department of Immunology, Albert Szent-Györgyi Medical School,
Faculty of Science and Informatics, University of Szeged

Institute of Biochemistry, Biological Research Centre

Szeged, 2024

The thesis is based on the following publications:

Bukva, M., Dobra, G., Gyukity-Sebestyen, E., Boroczky, T., Korsos, M. M., Meckes, D. G., Jr, Horvath, P., Buzas, K., & Harmati, M. (2023). Machine learning-based analysis of cancer cell-derived vesicular proteins revealed significant tumor-specificity and predictive potential of extracellular vesicles for cell invasion and proliferation - A meta-analysis. *Cell communication and signaling: CCS*, *21*(1), 333. https://doi.org/10.1186/s12964-023-01344-5. (**IF: 8.444, Q1**)

Bukva, M., Dobra, G., Gomez-Perez, J., Koos, K., Harmati, M., Gyukity-Sebestyen, E., Biro, T., Jenei, A., Kormondi, S., Horvath, P., Konya, Z., Klekner, A., & Buzas, K. (2021). Raman Spectral Signatures of Serum-Derived Extracellular Vesicle-Enriched Isolates May Support the Diagnosis of CNS Tumors. In Cancers (Vol. 13, Issue 6, p. 1407). MDPI AG. https://doi.org/10.3390/cancers13061407 **(IF: 6.575, Q1)**

Further publications:

Jakab, AE., Horváth, E., Molnár, D., **Bukva, M**., Bereczki. C., Validation of the Meditech ABPM-06 24-hour Blood Pressure Monitoring System in a Pediatric Population According to International Organization for Standardization Protocol 81060-2:2018. *Blood Pressure Monitoring* (**IF: 1.258, Q2)**

Welsh, JA., Goberdhan, DCI., O'Driscoll, L., Buzas, EI., Blenkiron, C., Bussolati, B., Cai, H., Di Vizio, D., Driedonks, TAP., Erdbrügger, U. et al. Minimal information for studies of extracellular vesicles (MISEV2023): from basic to advanced approaches. *Journal of Extracellular Vesicles* 2023. 10.1002/jev2.12404 B **(IF: 16.000, Q1) - Not included in the cumulative impact factor.**

Mezőlaki, N. E., Baltás, E., Ócsai, H. L., Varga, A., Korom, I., Varga, E., Németh, I. B., Kis, E. G., Varga, J., Kocsis, Á., Gyulai, R., **Bukva, M**., Kemény, L., & Oláh, J. (2024). Tumour regression predicts better response to interferon therapy in melanoma patients: a retrospective single centre study. *Melanoma Research*, *34*(1), 54–62. https://doi.org/10.1097/cmr.0000000000000935 **(IF: 2.166, Q1)**

Takács, A. T., **Bukva, M**., Bereczki, C., Burián, K., & Terhes, G. (2023). Diagnosis of Epstein-Barr and cytomegalovirus infections using decision trees: an effective way to avoid antibiotic overuse in paediatric tonsillopharyngitis. *BMC Pediatrics*, *23*(1). https://doi.org/10.1186/s12887-023-04103-0 **(IF: 2.386, Q1)**

Dobra, G., Gyukity-Sebestyén, E., **Bukva, M.**, Harmati, M., Nagy, V., Szabó, Z., Pankotai, T., Klekner, Á., & Buzás, K. (2023). MMP-9 as prognostic marker for brain tumours: A comparative study on serum-derived small extracellular vesicles. *Cancers*, *15*(3), 712. https://doi.org/10.3390/cancers15030712 **(IF: 5.207, Q1)**

Böröczky, T., Dobra, G., **Bukva, M.**, Gyukity-Sebestyén, E., Hunyadi-Gulyás, É., Darula, Z., Horváth, P., Buzás, K., & Harmati, M. (2023). Impact of experimental conditions on extracellular vesicles' proteome: A comparative study. *Life* (Basel, Switzerland), *13*(1), 206. https://doi.org/10.3390/life13010206 **(IF: 3.212, Q2)**

Takács, A. T., Bukva, M., Gavallér, G., Kapus, K., Rózsa, M., Bán-Gagyi, B., Sinkó, M., Szűcs, D., Terhes, G., & Bereczki, C. (2022). Epidemiology and clinical features of SARS-CoV-2 infection in hospitalized children across four waves in Hungary: A retrospective, comparative study from March 2020 to December 2021. *Health Science Reports*, *5*(6). https://doi.org/10.1002/hsr2.937 **(IF: 2.016, Q2)**

Ulbert, Á. B., **Bukva, M.**, Magyari, A., Túri, Z., Hajdú, E., Burián, K., & Terhes, G. (2022). Characteristics of hepatitis E viral infections in Hungary. *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, *155*(105250), 105250. https://doi.org/10.1016/j.jcv.2022.105250 **(IF: 8.809, Q1)**

Király, K., Váradi, O. A., Kis, L., Nagy, R., Elekes, G., Bukva, M., Tihanyi, B., Spekker, O., Marcsik, A., Molnár, E., Pálfi, G., & Bereczki, Z. (2022). New insights in the investigation of trepanations from the Carpathian Basin. *Archaeological and Anthropological Sciences*, *14*(4). https://doi.org/10.1007/s12520-022-01548-9 **(IF: 2.211, Q1)**

Fodor, E., Olmos Calvo, I., Kuten-Pella, O., Hamar, E., **Bukva, M.**, Madár, Á., Hornyák, I., Hinsenkamp, A., Hetényi, R., Földes, F., Brigitta, Z., Jakab, F., Kemenesi, G., & Lacza, Z. (2022). Comparison of immune activation of the COVID vaccines: ChAdOx1, BNT162b2, mRNA-1273, BBIBP-CorV, and Gam-COVID-Vac from serological human samples in Hungary showed higher protection after mRNA-based immunization [JB]. *European Review for Medical and Pharmacological Sciences*, *26*(14), 5297–5306. https://doi.org/10.26355/eurrev_202207_29321 **(IF: 3.252, Q2)**

Harmati, M., **Bukva, M.**, Böröczky, T., Buzás, K., & Gyukity-Sebestyén, E. (2021). The role of the metabolite cargo of extracellular vesicles in tumor progression. *Cancer Metastasis Reviews*, *40*(4), 1203–1221. https://doi.org/10.1007/s10555-021-10014-2 **(IF: 9.237, Q1)**

Dobra, G., Bukva, M., Szabo, Z., Bruszel, B., Harmati, M., Gyukity-Sebestyen, E., Jenei, A., Szucs, M., Horvath, P., Biro, T., Klekner, A., & Buzas, K. (2020). Small extracellular vesicles isolated from serum may serve as signal-enhancers for the monitoring of CNS tumors. *International Journal of Molecular Sciences*, *21*(15), 5359. https://doi.org/10.3390/ijms21155359 **(IF: 5.924, Q1)**

Cumulative IF: 60.697

Table of content

ABBREVIATIONS

Abbreviation	Meaning
ANOVA	Analysis of Variance
AUC	Area under the curve
BM	Brain metastsis of non-small cell lung carcinoma
C	Cost strength parameter
CA	Classification accuracy
CTRL	Lumbar disc herniation
DPBS	Dulbecco's Phosphate-Buffered Saline
ECM	Extracellular matrix
EV	Extracellular vesicle
FDR	False Discovery Rate
FPKM	Fragment per kilobase per million
GBM	Glioblastoma multiforme
HPA	Human Protein Atlas
IP	Immunprecipitation
LASSO	Least Absolute Shrinkage and Selection Operator
M	Meningioma
MISEV	Minimal information for studies of extracellular vesicles
NCBI	National Center for Biotechnology Information
NCI-60	National Cancer Institute 60
NSCLC	Non-small cell lung cancer
PC	Principal component
PCA	Principal Component Analysis
R^2	Coefficient of Determination
ROC	Receiver Operating Characteristic
RV	RV coefficient
SEQ	Size exclusion chromatography
sEV	Small extracellular vesicle
SNV	Standard Normal Variate
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TEM	Transmission electron microscopy
t-SNE	t-distributed Stochastic Neighbor Embedding
UC	Ultracentfiguation

1. INTRODUCTION

The present thesis addresses the diagnostic and prognostic capabilities of extracellular vesicles across various types of tumors. In its ***Introduction***, the thesis provides a concise overview of the theoretical framework that underpins this multifaceted study. This framework covers several key areas: extracellular vesicles, their role in tumor processes, their importance in biomarker research, the potential of Raman spectroscopy in tumor diagnostics, and the benefits of machine learning in cancer research.

1.1. Biological properties of extracellular vesicles

In recent decades, extracellular vesicles (EVs) have emerged as crucial mediators of intercellular communication in both eukaryotes and prokaryotes[1].

EVs are lipid bilayered particles that are secreted by all types of living cells into the extracellular space. Based on their biogenesis pathway, EVs could be classified into two basic types, namely exosomes and ectosomes[1,2].

Exosomes are of endosomal origin. Their biogenesis is initiated by the invagination of endosomal membranes, leading to the formation of multivesicular bodies enclosing intraluminal vesicles. Subsequently, these multivesicular bodies fuse with the cell's plasma membrane, thereby releasing enclosed vesicles, which can then be considered as exosomes in the extracellular milieu[1,2].

The other basic route of EV biogenesis is the release of plasma membrane-derived EVs (known as ectosomes). During this process, the plasma membrane undergoes dynamic alterations, leading to the budding and subsequent pinching off of vesicles containing specific cargo. In this way, among other particles, microvesicles, migrasomes, or ciliary ectosomes are secreted[1–3].

However, definitive molecular markers of the different biogenesis routes are not yet available, instead operational terms have been suggested to distinguish EV types based on their biophysical (e.g. density, size) or biochemical properties (present marker). Based on size, EVs could be classified into small EVs (sEV) (50-200 nm), which are the most abundant in the extracellular space, as well as medium-sized 200-1000 nm) and large (diameter ≥1 μm) EVs, which are found in lower concentrations[2].

This thesis follows the terminology used in the cited research when referring to EVs. In our own research, we utilized size-based classification.

Even though EVs are highly heterogeneous, they share a common molecular composition: they contain lipids, proteins, nucleic acids, and low molecular weight metabolites (**Figure 1**)[1,2].

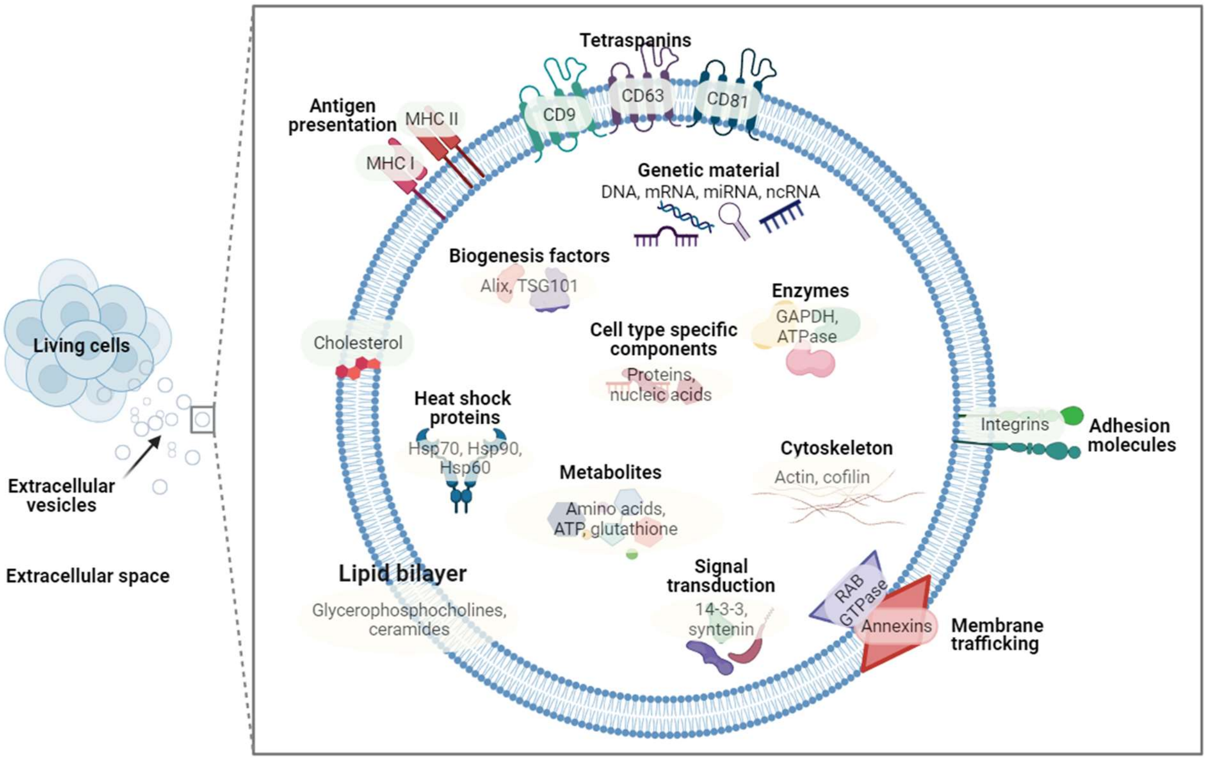


Figure 1. ***General molecular cargo of extracellular vesicles.*** All living cells release various types of extracellular vesicles (EVs) through different biogenesis pathways into the extracellular space. The typical EV is characterized by a composite structure comprising genetic materials, proteins, lipids and low-molecular-weight metabolites. Additionally, EVs possess unique cell-type specific components that vary according to the originating cell's nature.

The double lipid layer of EVs is primarily comprised of glycerophosphoethanolamines, phosphocholines, ceramides, glycosphingolipids, sphingomyelins, and cholesterol[4].

The cargo of EVs consistently includes transmembrane and lipid-bound proteins, such as tetraspanins, integrins, cell adhesion molecules, growth hormone receptors, and heterotrimeric G proteins[5]. Additionally, proteins present in the cytosol with capabilities for membrane or receptor binding are also present. These include factors necessary for the biogenesis of EVs, as well as proteins responsible for signal transduction or scaffolding functions[5].

EVs also carry cell-type-specific components, which are reflective of the characteristics of their producing cells. This specificity is illustrated through various examples: EVs from tumor cells are known to contain tumor antigens, indicative of their origin from malignant cells[6]. Exosomes derived from platelets include coagulation factors, aligning with the platelets' role in blood clotting[7]. Similarly, exosomes from dendritic cells are characterized by the expression of MHC-II-peptide complexes, crucial for immune response functions[5,8].

Numerous studies have revealed the significant role of EVs in maintaining cellular homeostasis and integrity by counteracting stress conditions[9].

Regardless of their heterogeneity, the overarching significance of EVs rests on their capability to transfer information from the donor cells – that produce them – to the recipient cells, that either internalize or interact with these EVs. This interaction plays a pivotal role in modulating the functions of the host cells, thereby regulating a wide range of physiological and pathological processes[10].

Research has demonstrated that EVs play a key role in various biological functions, including pregnancy and early development (notably in trophoblast implantation)[11], immune responses (encompassing both activation and suppression)[3], neuronal activities (such as axon growth and differentiation)[12], and processes related to angiogenesis and coagulation[7].

The impact of EVs is not limited to physiological processes, but also extends to pathological conditions, with a particular emphasis on tumor-related processes, which will be further discussed.

1.2. Extracellular vesicles in cancer

EVs, with their diverse molecular contents, significantly impact both physiological and pathological processes. Their role in cancer development and progression is particularly notable, representing a major area of research within oncology[13].

Several studies have highlighted the crucial role of EVs in enabling communication between cells in the primary tumor microenvironment.

Al-Nedawi et al. (2008) provided insights into how tumor-secreted vesicles transfer oncogenic molecules within a primary tumor[14]. Their study revealed that glioma cells expressing EGFRvIII release microvesicles containing this receptor, transferring it to neighboring EGFRvIII-negative cancer cells. The EGFRvIII mutation, prevalent in 25-64% of glioblastoma patients, is known for its continuous signaling in the absence of a known ligand. The reception of EGFRvIII through EVs initiates essential signaling pathways such as MAPK and PKB/Akt, fostering intensive cancer growth[14].

In addition to transferring oncogenic molecules, they also interact with stromal and endothelial cells. Microvesicles from glioblastoma, which carry mRNA, miRNA and angiogenic proteins, not only boost primary tumor growth but also promote endothelial cell proliferation[15]. Similarly, pancreatic cancer-derived EVs, carrying TSPAN8, stimulate angiogenic gene expression in endothelial cells[16].

EVs also participate in bi-directional communication between primary tumors and stromal cells. Exosomes from breast cancer-associated fibroblasts, for example, enhance the mobility and invasion of breast cancer cells via the Wnt-planar cell polarity pathway[17].

As cancer progresses, tumor-derived EVs dynamically affect the cellular and molecular environment of the stroma, including the extracellular matrix, which is often associated with increased tumor invasiveness. For example, matrix metalloproteinase (MMP)-containing EVs initiate proteolysis in the ECM, which gives way to tumor cell migration and angiogenesis[18].

EVs that carry ECM elements, such as fibronectin, promote cell mobility by aiding in adhesion assembly[19]. Proteomic analysis has identified crucial molecules including annexins, ITGA3, and ADAM10 in exosomes, which are associated with local invasion and cell migration[20].

Additionally, EVs are implicated in the epithelial-mesenchymal transition (EMT). Overexpression of HRAS in cancer cells leads to exosomes being packed with mesenchymal markers, potentially inducing EMT in recipient cells[21]. Exosomes from the metastatic breast cancer cell line MDA-MB-231, when stimulated with linoleic acid, induce an EMT-like process in epithelial MCF10A cells[22].

Given their diverse molecular contents and crucial role in cell-cell communication, particularly in cancer progression and metastasis, EVs are regarded as a potentially rich source of biomarkers, offering insights into oncogenic processes and tumor microenvironment dynamics[23].

1.3. Extracellular vesicles in biomarker research

Recent clinical research has highlighted that EVs could serve as novel tools for various therapeutic approaches, including oncotherapy, vaccination, immune-modulatory or regenerative therapies, and drug delivery[24]. Nevertheless, the key role of EVs in liquid biopsy-based biomarker research stands out among these potential applications[24,25].

Previous research has not only elucidated that the concentration of EVs in the blood of cancer patients is significantly higher on average[26], but also that these EVs originate not only from the tumor tissue but also as a result of immune processes associated with cancerous disease[27].

Consequently, it seems intuitive to consider EVs isolated from serum or plasma samples as a potential source of biomarkers that can provide information on many aspects of tumor disease through a non-invasive way[25] (**Figure 2).**

Despite the fact that EV-based biomarker research has been ongoing since 1993[28], it has long been an unanswered question whether it is really beneficial to isolate EVs from patients' serum samples or whether the same clinical efficacy can be achieved by simply analyzing the whole serum (without EV isolation).

Our research group was the first to shed light on this quandary in a previously published study, showing that isolating sEVs from serum significantly improves the signal-to-noise ratio[29].

In said study, we aimed to distinguish groups of patients with glioblastoma, meningioma, and brain metastases of non-small cell lung cancer (NSCLC), as well as from a control group with intervertebral disc hernia.

Comparison of the proteomes of the two sample types – whole serum and sEV samples – showed that the sEV isolation procedure significantly decreased the abundance of apolipoproteins and increased the concentration of non-tissue-specific EV markers (ITGA2B, ITGB3, LGALS3BP), epithelial (CD5L) and platelet EV markers (STOM, TSPAN9), and significantly altered the intensity of hundreds of protein levels.

Using the same data analysis pipeline on the proteome of whole serum and sEV samples, it was found that the protein content of EVs is more suitable to discriminate between patient groups.

The findings suggested that – although it is not possible to eliminate contaminations in the sEV samples – the sEVs do carry tumor-specific patterns in a relatively more concentrated form compared to the whole serum, in which signals may be masked by the abundant serum proteins and lipoproteins[29,30].

Carrying this concentrated tumor-specific molecular information, EVs are able to enter the circulation by crossing different biological barriers (e.g., blood-brain-barrier)[31], from where they can be isolated by various methods and subjected to downstream analysis.

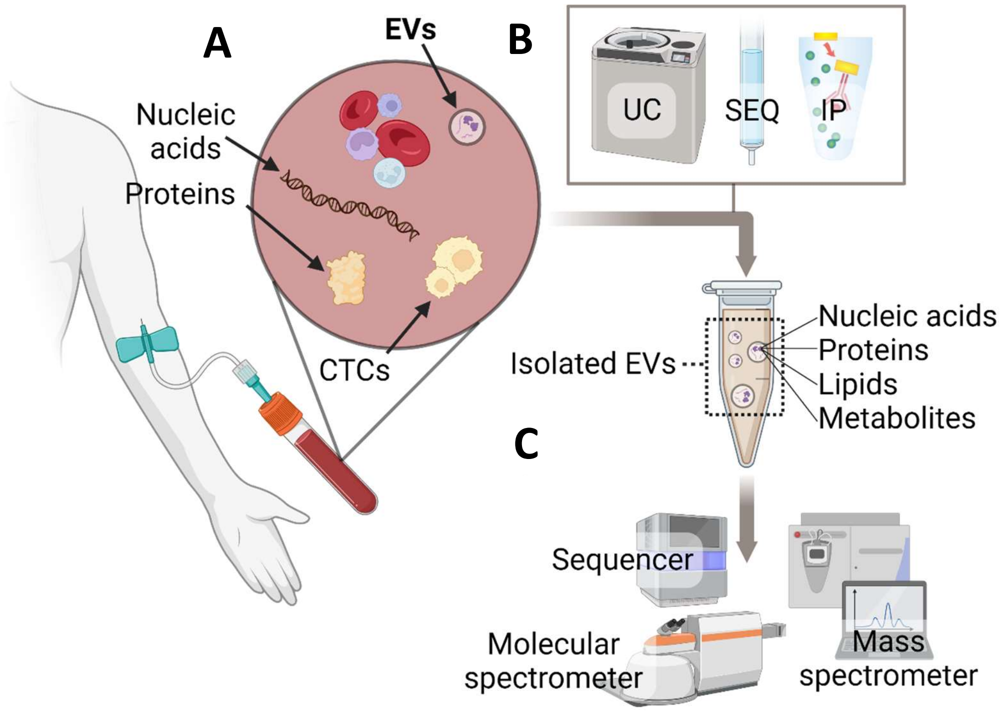


Figure 2. ***Extracellular vesicles in liquid biopsy-based biomarker research***. (**A**) During the course of a cancerous disease, tumor tissue and tumor-associated immune processes release molecular information into the blood in the form of RNA, DNA, proteins, tumor cells and extracellular vesicles (EVs). The EVs can be isolated from the collected blood samples (especially, from the serum or plasma) using various techniques, including differential centrifugation, size-exclusion chromatography, or immunoprecipitation (**B**). Subsequently, the nucleic acid content of the sample enriched in EVs can be analyzed by sequencing methods, their lipid, metabolite and protein content by mass spectrometry, and all these comprehensively by molecular spectroscopy, such as Raman spectroscopy (**C**). (CTC: circulating tumor cells; EV: extracellular vesicles; IP: immunoprecipitation; SEQ: size exclusion chromatography; UC: ultracentrifugation).

1.4. Approaches in biomarker research

As previously discussed, the investigation of EVs in the identification of biomarkers holds significant potential due to their ability to be enriched with the tumor-specific molecular composition[25,29].

Currently, in EV-based biomarker research, we can distinguish two approaches: (I) on one hand, individual molecule types can be examined separately using various omics methods, and (II) on the other hand, the full molecular complexity of

EVs can be encompassed, characterizing them as a whole, for example, through vibrational spectroscopy methods[32].

The first category covers those methods that are used to analyze either only lipids or only proteins, nucleic acids, metabolites, etc. These methods are generally well standardized and there are many examples of their effectiveness.

For example, Menck and colleagues demonstrated that the combined analysis of EMMPRIN$^+$, MUC1$^+$, EGFR$^+$, and EpCAM$^+$ EVs effectively distinguishes between the tumor and control groups[33].

Similarly, Jakobsen and colleagues, through proteomic analysis of EVs isolated from the plasma of NSCLC patients, identified 30 proteins. The combined examination of this panel enabled them to distinguish cancer patients from healthy controls with high accuracy[34].

In addition to proteins, miRNAs can also be potential targets of significant interest. A newly emerging meta-analysis, which reviewed 2,395 articles, highlighted that in the diagnosis of prostate tumors, miR-141 and miR-221 could play a crucial role due to their high sensitivity and specificity[35].

Beyond these examples, there are numerous other studies that examine the diagnostic and prognostic capabilities of EVs in the context of melanoma[36], breast cancer[37], or colorectal cancer[38].

As demonstrated by the aforementioned examples, various successful research studies have been conducted using well-known and standardized omics techniques. However, the different omics data sets do not overlap to a large extent, and measurements obtained with one omics approach often do not correlate well with data obtained with other methods. Thus, it is likely that different omics approaches assess different parts of the complex pathophysiology of disease development and progression, and analysis of a single omics subset provides a biased and incomplete picture of the underlying biology[39].

This biased picture is particularly challenging in the search for diagnostic and monitoring biomarkers in malignancies with pronounced inter-individual variability, such as glioblastoma multiforme. In such cases, the complex and varied molecular landscape between patients makes the identification of reliable and consistent biomarkers significantly more difficult[40]. Restriction to a limited number of proteins, nucleic acid or lipid biomarkers leads to a lack of generalizability and diagnostic value;

thus, what works well in the cohort used to identify biomarkers may not work for new patients[41–43].

The second category means examining the entire molecular composition of tumor-derived EVs. This is an emerging and promising approach to navigate the challenges in biomarker discovery for highly heterogeneous cancers. Vibrational spectroscopy techniques – like Raman spectroscopy – offer a novel, alternative approach to omics methods[44].

Raman spectroscopy is a non-destructive analytical technique that does not require any labeling of the sample. It employs the inelastic scattering of laser light to investigate the vibrational characteristics of molecules. When laser light interacts with a sample, most photons are elastically scattered (Rayleigh scattering) without changing energy. However, a small fraction of the light is inelastically scattered, experiencing a shift in energy due to the interaction with the vibrational states of the molecules in the sample. This energy shift, observed in the scattered light, corresponds to the vibrational energy of the chemical bonds within the molecules of the sample[45]. By analyzing these shifts, Raman spectroscopy provides a molecular fingerprint of the sample. This fingerprint is directly related to the chemical composition and molecular structure of the sample, making it a powerful tool for chemical analysis and identification of substances[44].

Consequently, Raman spectroscopy can generate a detailed spectral signature, representing the entire chemical composition of a sample, thereby eliminating the need to identify specific types of proteins, nucleic acids, or lipids for biomarker identification[46].

Recent research indicates that the application of Raman spectroscopy for analyzing the complete molecular structure of diverse sample types holds potential for developing innovative diagnostic approaches in clinical settings[47–49].

In vitro experiments have particularly highlighted the diagnostic potential of this technique, especially when applied to EVs. These studies have demonstrated remarkable diagnostic efficacy. For instance, the work of Parks et al. showcased the ability of Raman spectroscopy to differentiate between EVs emitted by lung cancer cells and those from healthy cells, achieving a sensitivity of 95.3% and a specificity of 97.3%. This differentiation was further enhanced through the use of principal component analysis to identify unique spectral variations[50]. Another research by Charmichael et al. showed that EVs derived from pancreatic cancer cells could be reliably distinguished from those emanating from normal pancreatic epithelial cells, with an accuracy of 90%[51].

To summarize the above, the two approaches inherently serve different purposes. The application of omics methods provides us the opportunity to identify specific molecules, uncover particular pathways, and strategies characteristic of tumors. However, these methods might encounter challenges in diagnosing tumors that present diagnostic difficulties, such as CNS malignancies[40]. On the other hand, vibrational spectroscopy methods can overcome these challenges, but due to the complexity of biological samples, it is much less likely to specifically identify individual molecules. Nevertheless, because these methods encompass the entire molecular complexity, they may offer greater diagnostic value[50,51].

As demonstrated previously, EVs can be seen as a rich source of biomarkers across various approaches. However, as methods suitable for unraveling the complex molecular content advance, there is an increasing need for more sophisticated data analysis procedures to complement them[52]. These advanced techniques are essential for accurately deciphering complex patterns within biological data, which brings us to the next chapter, which focuses on the vital role of sophisticated statistical modeling and machine learning in cancer research, particularly in the search for efficient and reliable biomarkers.

1.5. The increasing prevalence of machine learning in cancer research

Machine learning, as part of artificial intelligence, involves sophisticated algorithms that semi- or fully automate problem-solving processes. These algorithms, grounded in mathematical and statistical principles, progressively enhance their performance by emulating the concept of 'learning.' As they are exposed to more data, these systems adapt and refine their decision-making capabilities, thereby solving a wide array of complex problems with increasing efficiency and accuracy[53].

This innovation has significantly revolutionized many fields, particularly in biomedical research, where machine learning techniques have the unprecedented ability to capture the tremendous complexity inherent in biological “big data”[54].

The impact of machine learning is particularly pronounced in cancer research, which has seen an exponential growth in data volume and complexity over the past 30 years, primarily due to advancements in high-throughput sequencing, omics technologies, and single-cell analysis[55]. Pioneering projects like The Cancer Genome Atlas (TCGA) and the Human Protein Atlas (HPA) exemplify this surge, housing vast amounts of data crucial for understanding cancer at a molecular level[55].

The overarching goal of extensive data collection in cancer research is to comprehensively analyze the characteristics of malignancy. This involves unraveling the intricate relationships between diverse genomic, proteomic, transcriptomic, and metabolomic datasets, and assessing their impact on clinical outcomes like disease incidence, overall survival, progression-free survival, and therapeutic response[56].

Conventional statistical approaches often struggle with this high dimensionality and the broad range of data types. Their limitations become apparent in capturing the nuanced relationships present in diverse biomedical data sets[57].

Typically, conventional statistical methods follow a top-down approach where there are prior assumptions about the relationship between the variables (such as in linear and logistic regression)[57,58] (**Figure 3**).

This assumption makes the interpretation of the results straightforward and the relationships between variables easy to comprehend[59]. For example, we have preliminary assumptions that high MMP-9 blood concentrations and overall survival are negatively correlated.

In light of this, we expect that we can make a good estimate of someone's overall survival time (which in this case is the outcome variable) from their MMP-9 concentration (which is the predictor variable).

However, this approach is limited by the fact that the link between input and output is user-selected and may result in a less accurate prediction model if the actual input-output relationship is not well-represented by the chosen model. This can happen when a linear regression is chosen by the user, but the relationship between input and output is non-linear or when many input variables are involved.

For example, many clinical characteristics and risk factors exhibit non-standard distributions or nonlinear relationships[60]. For example, certain risk factors may have a U-shaped relationship with a disease outcome, where both low and high levels of the risk factor are associated with increased risk. Conventional statistical models, which assume standard distributions (e.g., normal distribution) and linear connections, may not adequately capture these non-standard distributions and nonlinear relationships. Consequently, these models might overlook important associations and fail to detect complex patterns or interactions within the data[61].

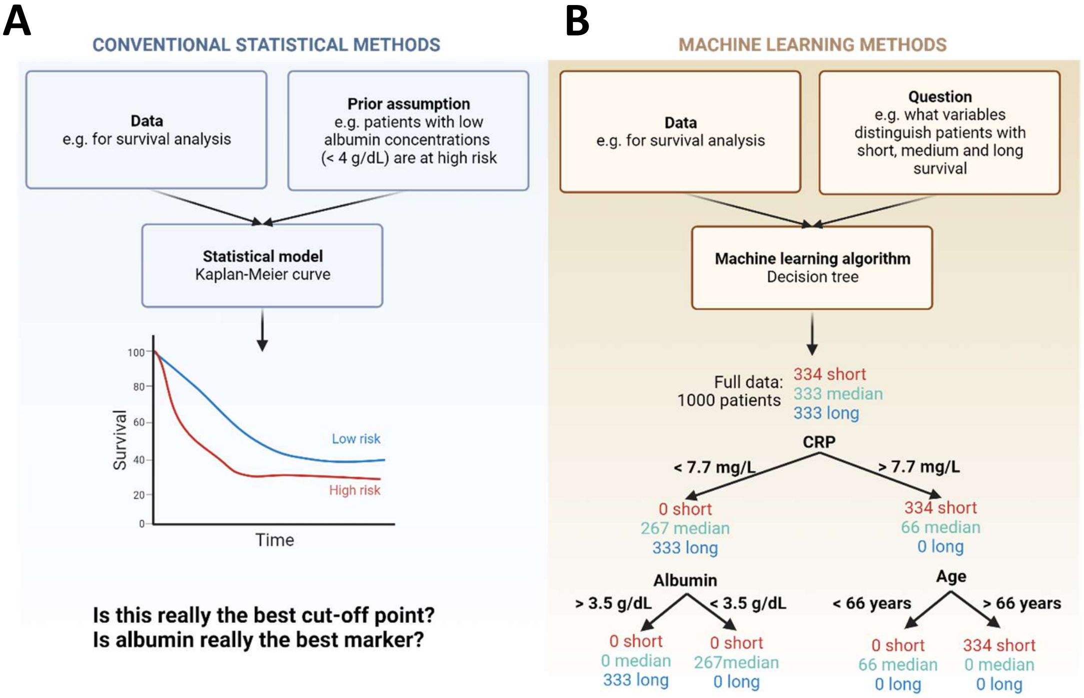


Figure 3. ***Differences between conventional statistical and machine learning methods***. The conventional statistical method (**A**), employing a top-down approach, presupposes the existence of a pre-selected variable with an established cut-off value, which we aim to use for generating Kaplan-Meier curves. There is a discernible difference in the survival of groups above and below this cut-off value, but this raises the question of whether this pre-chosen cut-off is indeed the optimal threshold. In addition, other variables could be considered in addition to albumin. In contrast, in machine learning analysis (**B**), which utilizes a bottom-up approach, we define our groups and create categories with low, medium, and long survival. By employing a decision tree algorithm and conducting a series of automatic calculations, we can identify the variables, their combinations, and cut-off values that most effectively differentiate the three groups. (CRP: C-reactive protein.)

In contrast, machine learning follows a "bottom-up" approach. Throughout the process, there are no prior assumptions about the model that describes the relationship between variables. Instead, an algorithm autonomously uncovers the intricate relationships among variables within a complex dataset[59,62].

For instance, when aiming to estimate the overall survival of distinct patients while having access to variables such as age, blood MMP-9 concentrations, sentinel positivity and socioeconomic status, method within the realm of machine learning, such as the Random Forest, can be employed to automatically build decision trees describing the connection between the predictor variables and the patient survival[63].

In summary, machine learning approaches, devoid of prior assumptions, are adept at unraveling complex relationship systems within large datasets, a capability increasingly essential in modern cancer, especially, in biomarker research[64]. Their application can be particularly beneficial on large data sets such as proteomic or Raman spectral data.

1.6. Research purpose and significance

Numerous studies have highlighted the role of EVs in tumorous processes, leading to efforts to include them in liquid biopsy based diagnostic methods[25]. The majority of these studies have demonstrated that the analysis of EVs can be used to differentiate between tumorous and control samples or to subcategorize tumor types based on their properties (e.g. chemosensitivity)[33–36,38,43,65–67].

However, there are still several unexplored areas regarding the potential utility of EVs. For instance, it is still under exploration whether the molecular composition of EVs can predict the invasion capacity or proliferation rate of the donor cells, or whether they could provide information on tumor-specific signaling pathways or strategies. Furthermore, as most of the studies investigate a limited number of patient groups, the degree of specificity of the molecular pattern carried by EVs of different tumor types is not fully elucidated.

While there are notable instances of Raman spectroscopy of EVs demonstrating remarkable diagnostic efficacy, the effectiveness of this technique in analyzing highly heterogeneous tumor types or those presenting significant diagnostic challenges, such as CNS tumors, remains less clear.

With this in mind, the present thesis summarizes the findings of two studies. The first part focuses on the specificity and potential clinical utility of the proteome of different tumor-derived EVs, analyzed through a meta-analysis of *in vitro* data. The second part, as a clinical study, assesses the effectiveness of using Raman spectroscopy to analyze serum-derived sEVs for diagnosing CNS tumors. Common feature of the two studies is that, in addition to conventional statistical methods, they rely heavily on machine learning methods to process data from EVs. Evaluating results using machine learning methods helps to fully exploit the potential of the disease-related molecular composition of EVs.

2. AIMS

The primary aim of this thesis is to highlight the potential of the tumor-associated molecular content carried by EVs. This includes the potential role of EVs in tumor diagnostics, differential diagnosis, prognosis, and drug targeting through a better understanding of cancer characteristics.

For this purpose, the thesis summarizes the results of two studies, a meta-analysis on in vitro data and a clinical study on clinical serum samples.

In the ***meta-analysis*** of the proteome of EVs isolated from the supernatants of 60 different cell lines from nine tumor types (NCI-60 panel), with the following aims:

1. To assess the degree of tumor specificity of the total proteome and proteins common to all 60 EV samples, and to select proteins that most effectively discriminate between the nine tumor types.
2. Elucidate the biological functions of the discriminative proteins for tumor characteristic patterns.
3. Select the proteins that can predict the invasion capacity and proliferation time of donor cells.

In the ***clinical study*** to analyze the Raman spectra of sEV-enriched isolates from serum samples of patients with glioblastoma multiforme (GBM), brain metastasis (BM), meningioma (M), and lumbar disc herniation (CTRL), we aimed to:

4. To build a classification model capable of distinguishing between different patient groups based on the Raman spectra of the sEV-enriched isolates.

3. MATERIALS AND METHODS

3.1. Meta-analysis of in vitro data

The proteomic dataset foundational to the meta-analysis was provided by Hurwitz and colleagues[68]. In their research, vesicles isolated from cell line supernatants were referred to as "EVs". Consequently, this thesis adopts the term "EV" in discussing findings connected to this dataset.

3.1.1. Cell lines

EVs were isolated from the supernatants of NCI-60 (National Cancer Institute 60) cell lines. This panel contains 60 different cell lines from nine tumor types (breast, CNS, colon, kidney, leukemia, lung, melanoma, ovarian, prostate[69–72]).

3.1.2. Proteomic data

We obtained the proteomic data of EVs as freely downloadable supplementary material[68]. This data set contains the spectral count and intensity of 6,701 proteins for 60 EV isolates harvested from the NCI-60 panel. In our study, we used the intensity values for the analyses. Before the analyses, the intensities were logarithmized ($\log_2$) to increase the linearity and reduce the variance. Imputation of missing values was not performed, as the 0 values in the data matrix used do not represent missing values, but the absence of proteins in the EV isolate.

3.1.3. Data on the invasion capacity of NCI-60 cell lines

The invasion phenotype of the 60 cell lines were obtained from the publication of DeLosh et al. as freely downloadable supplementary material[73].

Briefly, DeLosh et al. utilized CIM (cellular invasion/migration)-Plate 16 to determine the invasion capacity of the NCI-60 panel. The CIM Plate-16 consists of two chambers, one below the other. The chambers are separated by a microporous membrane. Microelectronic sensors are integrated at the bottom of the pores in the lower chamber on the other side of the membrane. The migration of cells from the upper chamber to the lower chamber in response to a chemoattractant leads to their interaction and attachment to the electrical sensors, hence causing an elevation in impedance. The impedance correlates to increasing numbers of migrated cells on the underside of the membrane, and cell index values reflecting impedance changes are automatically and continuously recorded by the e Roche xCELLigence Real-Time Cell

Analyzer DP instrument. Therefore, cell migration activity can be monitored via the cell index profile.

The invasion phenotype of 60 cell lines was determined by plotting the cell index (reflecting the mass of the cell detected) as a function of analysis time and then calculating the area under the curve (AUC). We used the average AUC for each cell line as published in the original article, but refer to it as invasion capacity for ease of interpretation.

3.1.4. Data on the proliferation of NCI-60 cell lines

Doubling time of NCI-60 cell lines data were obtained from the National Cancer Institute website although[74], to facilitate interpretation, we refer to it as proliferative capacity for ease of interpretation.

3.1.5. Data on RNA expression of the NCI-60 cell lines

Microarray gene expression data was downloaded from the NCBI Gene Expression Omnibus (accession number: GSE32474)

3.1.6. Data on the in situ tissue expression and survival

In our study, we obtained information from the Human Protein Atlas database regarding the *ex vivo* tissue expression of specific proteins and the overall survival time (in years) of patients corresponding to the tissue samples[75].

3.1.7. Classification of EV samples

During the classification, we attempted to classify the 60 EV samples into their respective nine tumor types (breast, central nervous system—CNS, colon, kidney, leukemia, lung, melanoma, ovary, prostate).

We applied Multivariate logistic regression on the proteomic data set for classification purposes. First, the 60 EV sample was classified based on shared proteins and then on the entire proteome. After classifying based on the entire proteome, we aimed to identify a discriminant protein panel for the nine tumor types.

In every classification procedure, the data set was split 50–50%, creating a Train and a Test set. We utilized the Least Absolute Shrinkage and Selection Operator (LASSO) method to score the proteins on the Train set according to their importance in distinguishing the tumor types (this score is the regression coefficients). This value can be negative, positive, or zero, suggesting a negative or positive effect on the probability of classifying into a certain tumor type, or an irrelevant protein.

In LASSO, the so-called cost strength parameter (C), which can vary from 0.001 to 1000, indicates how strict the scoring is (affecting the number of proteins scored as irrelevant/meaningless). In this study, this value was set to 1, which resulted in neither too strong nor too weak scoring, and allowed us to select characteristic proteins for each of the nine tumor types. The optimal value of the parameter C was determined by five fold cross-validation of the train set and fixed at the point where the highest classification efficiency was measured.

The list of characteristic proteins (discriminative protein panel) for the nine tumor types included only proteins selected by the LASSO algorithm. Classification was again performed on the Test data set based on the proteins selected.

The efficiency of the classification was given by classification accuracy (number of correctly classified samples divided by the total number of samples). The success of the classification was visualized using confusion matrices.

Orange 3.27.0 software was used to conduct the classification and create figures[76].

3.1.8. Regression for invasion and proliferative capacity

Multivariate linear regression was performed with LASSO (parameter C = 1) to predict invasion and proliferation capacity. LASSO played the same role in regression as it did in classification.

It is important to note that the approach (CIM Plate-16) used to determine invasiveness of the cell lines has only been shown to be applicable to solid tumors [44]. Therefore, leukemia was not included in the determination of proteins predictive of invasion capacity.

The procedure involved splitting the data into a Train and Test set in a 50-50% ratio. The Train set was used to identify proteins that could potentially predict invasion and proliferation capacity using LASSO. The relationship between the selected proteins and invasion/proliferation capacity was then investigated using the Test set through multivariate linear regression.

A significance level of $p < 0.05$ was used. The efficiency of the regression was measured by the coefficient of determination (R^2). Multivariate linear regression and visualization were performed using Orange 3.27.0 and GraphPad Prism 8.4.3 (San Diego, CA, USA).

3.1.9. Pathway enrichment analysis

We utilized ShinyGO 0.76.3 for Gene Ontology Enrichment Analysis to determine the biological processes, molecular functions, and cellular components whose proteins are overrepresented in our data set[77]. The ShinyGO parameters were set to default.

Reactome (v82) was applied on the discriminative protein panel for simultaneous enrichment analysis of each sample in order to compare the 60 EV samples in terms of their associated signal pathways[78]. The Reactome parameters were set to default.

Value of $p < 0.05$ corrected with the False Discovery Rate (FDR) method was considered significant.

3.1.10. Hierarchical clustering

Hierarchical clustering based on proteins was performed after row centering and unit variance scaling. Both rows (proteins) and columns (EV samples) were clustered using correlation distance and complete linkage. Hierarchical clustering based on the Reactome results was performed on raw data, without any adjustment. The rows (pathways) were clustered using correlation distance and complete linkage.

Hierarchical clustering was performed using Morpheus software[79].

3.1.11. T-distributed Stochastic Neighbor Embedding

To visualize the proteomic data in a 2-dimensional space, we utilized the t-distributed Stochastic Neighbor Embedding (t-SNE) method. For t-SNE visualization, we used Orange 3.27.0.

3.1.12. Examining the similarity between the EV proteome and the cellular RNA profile

The similarity of protein and RNA profiles of EV samples and cells for each variable was tested by Spearman's correlation analysis, the results of which were plotted on heatmaps. In addition, the concordance of the two matrices (RNA profile of cells and protein content of EVs) was characterized overall with RV coefficients introduced by Escoufier[80].

In data analysis, the *RV* coefficient is a multivariate generalization of the squared correlation coefficient, depicting the similarity between two matrices of quantitative variables. The *RV* coefficient takes values between 0 and 1.

The analysis was performed using the *omicade4* package in the R statistical framework[81].

3.2. Clinical study on serum-derived extracellular vesicles with Raman spectroscopy

3.2.1. Patients

Blood samples of 138 patients treated at the Department of Neurosurgery at the University of Debrecen were analyzed. Samples were obtained from patients with glioblastoma multiforme (GBM), brain metastasis of non-small cell lung cancer (BM), meningioma (M). Patients with spinal disc herniation (a non-cancerous CNS disease) served as control CTRL (**Table 1**). Each patient signed an informed consent form. The study was conducted in accordance with the Declaration of Helsinki, and ethical approval was obtained from two independent bodies (51450-2/2015/EKU (0411/15), Medical Research Council, Scientific and Research Ethics Committee, Budapest, October 30, 2015 and 121/2019-SZTE, University of Szeged, Human Investigation Review Board, Albert Szent-Györgyi Clinical Centre, Szeged, 19 July 2019).

Table 1. ***Patient cohort.***

Patient	No. of	Age (years)			Sex	
Groups	Patients	Range	Mean	Median	Male (%)	Female (%)
CTRL	36	20–81	53.6	54	16 (44.4)	20 (55.6)
GBM	46	33–82	64.3	66	28 (60.9)	18 (39.1)
BM	28	42–82	63.5	62.6	18 (64.3)	10 (35.7)
M	28	30–79	58.6	60	5 (17.9)	23 (82.1)

3.2.2. Preparation of serum samples, sEV isolation and characterization

Briefly, after 1 hour of blood clotting at room temperature, sEV isolation from serum samples was performed via differential centrifugation (20 min at 3000× g, 10 °C; 30 min at 10,000× g, 4 °C; 70 min at 100,000× g, 4 °C). After the last centrifugation step, the pellet was resuspended in Dulbecco’s phosphate-buffered saline (DPBS) and was stored at −80 °C until further processing.

To characterize sEVs, we followed the main suggestions and requirements included in the guideline 'Minimal Information for Studies of Extracellular Vesicles 2018' (MISEV 2018)[2].

sEVs were diluted in particle-free DPBS and analyzed using a NanoSight NS300 instrument with 532 nm laser (Malvern Panalytical Ltd., Malvern, UK). Six videos of 60 s were recorded for each sample under constant settings (Camera level: 15; Threshold: 4, 25 °C; 60–80 particles/frame) and analyzed to obtain data on size distribution and particle concentration.

Classical EV markers were presented by Western blot analyses using NuPAGE reagents and an XCell SureLock Mini-Cell System (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's protocols. For detection of the CD81, Alix and Calnexin markers, we used rabbit anti-human CD81 (1:1000, Sigma-Aldrich, St. Louis, MO, USA), rabbit anti-human Alix (1:1000, Sigma-Aldrich, St. Louis, MO, USA) and rabbit anti-human Calnexin (1:10,000), Sigma-Aldrich, St. Louis, MO, USA) primary antibody and HRP-conjugated anti-rabbit IgG (1:1000, R&D Systems, Minneapolis, MN, USA) secondary antibody. THP-1 cell line (ATCC, Teddington, UK) lysate was used for positive control for Calnexin.

In order to examine sEV morphology, transmission electron microscopic (TEM) analysis was performed using a Tecnai G2 20 X-Twin type instrument (FEI, Hillsboro, OR, USA), operating at an acceleration voltage of 200 kV. For TEM measurements, the samples were dropped on a grid (carbon film with 200 Mesh copper grids (CF200-Cu, Electron Microscopy Sciences, Hatfield, PA, USA) and dried without staining or other fixation procedure.

3.2.3. Raman spectroscopy

The characterization of sEVs was performed utilizing Raman spectroscopy with a Senterra II Microscope (Bruker) operating in a backscattering configuration. The samples were subjected to centrifugation, air-dried at room temperature after being drop-cast onto a calcium fluoride substrate, and then analyzed using a 50× objective (Olympus) and a 532 nm excitation laser. A laser power of 12.5 mW and an integration time of 30 seconds (2 coadditions) were employed, with an interferometer resolution of 1.5 cm^{-1}. The resulting spectra were baseline-corrected and averaged (5 spectra per sample) using the OPUS software provided by Bruker, and the spectral range of 799.5 cm^{-1} to 3100.5 cm^{-1} was used for further analysis. The laser spot produced by the optical setup was approximately 15 μm, much smaller than the average size of the air-dried sample, allowing for fine-tuning of the sampling position to avoid duplication.

3.2.4. Data adjustment

The baseline-corrected data underwent row normalization through the application of the Standard Normal Variate (SNV) method. This method transformed the mean to 0 and the standard deviation to 1, thus rendering all spectra comparable in terms of intensity. PCA with unit variance scaling was applied to the SNV-normalized spectra in order to reduce the dimensions of the multivariate data. The original variables (wavenumbers) were transformed into a smaller number of new variables, known as the principal components (PCs).

The data adjustment was performed utilizing the Orange 3.27.0 software, developed in Ljubljana, Slovenia.

3.2.5. Classification

In order to construct and evaluate classification models, the four patient groups were compared in pairs (i.e., each patient group was compared with the control group, and GBM with the BM group).

The linear Support Vector Machine (SVM) algorithm was utilized for the classification of the samples, resulting in the generation of classification models for each compared group.

To begin with, the data was randomly partitioned into a Train and Test set in a 90-10% ratio. The SVM algorithm was employed on the Train set to identify a hyperplane that would distinguish the compared groups in the PCA-transformed space, thereby producing a trained SVM model.

The trained model then assigned group-membership scores (ranging from 0 to 1) to the Test samples based on their positions and distances from the separating hyperplane. In practice, the classification decisions were made based on the relative location of the test samples with respect to the hyperplane, which was expressed as their group-membership scores.

To make predictions about the test samples, a minimum score threshold was established. Test samples with scores higher than this threshold were classified into the target group of interest.

This train-test split procedure was repeated 100 times. The classification efficacy was evaluated in terms of sensitivity (the proportion of positive samples that were correctly identified), specificity (the proportion of negative samples that were correctly identified), classification accuracy (CA) and the AUC value obtained from the Receiver Operating Characteristic (ROC) analysis.

The Orange 3.27.0 and GraphPad Prism 8.4.3 (San Diego, CA, USA) software packages were employed for the purposes of classification and efficacy evaluation.

3.2.6. Determining the spectral differences

The correlation between the obtained principal components (PCs) and the different patient groups was evaluated using the FreeViz method[82]. The FreeViz approach presents the multivariate data in a 2-dimensional scatter plot, allowing the separation of samples from different patient groups.

Samples are represented as dots and PCs as vectors in the scatter plot. The FreeViz optimization of the display emphasized the significance of PCs in classifying the groups, where important PCs had longer vectors. The direction of PC vectors was also insightful, as it indicated which PCs were most useful for distinguishing between groups. When a region in the scatter plot was primarily populated by samples of a patient group, PC vectors in that direction could be seen as good indicators of group membership. The more a PC vector approached perpendicularity to the imaginary line separating the groups, the more it was effective in distinguishing them.

The statistical differences between PCs were analyzed using Welch's t-test. Since PCs are linear combinations of the original variables (wavenumbers), it was possible to determine the wavenumbers with the most impact on a given PC, thus separating the groups compared.

Results with a p value of less than 0.05 were considered significant. The FreeViz analysis was performed using the Orange 3.78.0 software.

4. RESULTS

4.1. Findings of the meta-analysis

4.1.1. Shared proteins of EVs are related to EV biogenesis processes

The proteomic data set in our meta-analysis, encompassing 60 EV samples, revealed a total of 6,071 proteins. Of these, 5,908 proteins were quantified in terms of intensity and are collectively referred to as the entire proteome in our research.

Gene Ontology Enrichment Analysis linked this entire proteome to a wide array of biological processes, molecular functions, and cellular structures. Notable among these are neutrophil-mediated immunity, cell adhesion to the extracellular matrix, and secretory vesicles and granules (**Supplementary table 1**). The fold enrichment values, reflecting the extent of gene overrepresentation in certain pathways, varied between 1.68 and 3.01. This suggests a significant discovery of proteins, at least 1.68 times higher than expected by chance, from the specified signal pathways.

Furthermore, our analysis identified 213 proteins consistently present across all EV samples, defining what we term the core proteome. Examination of this core proteome highlighted its involvement in pathways related to both intracellular and extracellular vesicle biogenesis, including co-translational protein targeting to the membrane, RNA binding, and cytosolic ribosomes (**Supplementary table 2**). Notably, the core proteome showed a more pronounced association with each biological pathway compared to the complete proteome, as indicated by higher fold enrichment values ranging from 3.78 to 33.12.

4.1.2. Entire proteome of EVs resulted higher classification accuracy of tumor cells lines than core proteome

In our research, we first turned our attention to the core proteome, analyzing it for tumor-specific patterns using a logistic regression classification model. Interestingly, this smaller portion of the entire proteome, though impacting only a limited number of biological processes, proved sufficiently informative to distinguish to a degree between various tumor types, such as kidney, lung, leukemia, and melanoma, as depicted in **Figure 3A, C**. The model's classification accuracy reached 49.14%, significantly surpassing the 11.1% accuracy expected from random classification.

Subsequent analysis with a one-way ANOVA showed that the tumor type influenced the average intensity of the core proteome ($p < 0.0001$). However, Pearson's correlation analysis indicated that this difference was not attributable to variations in

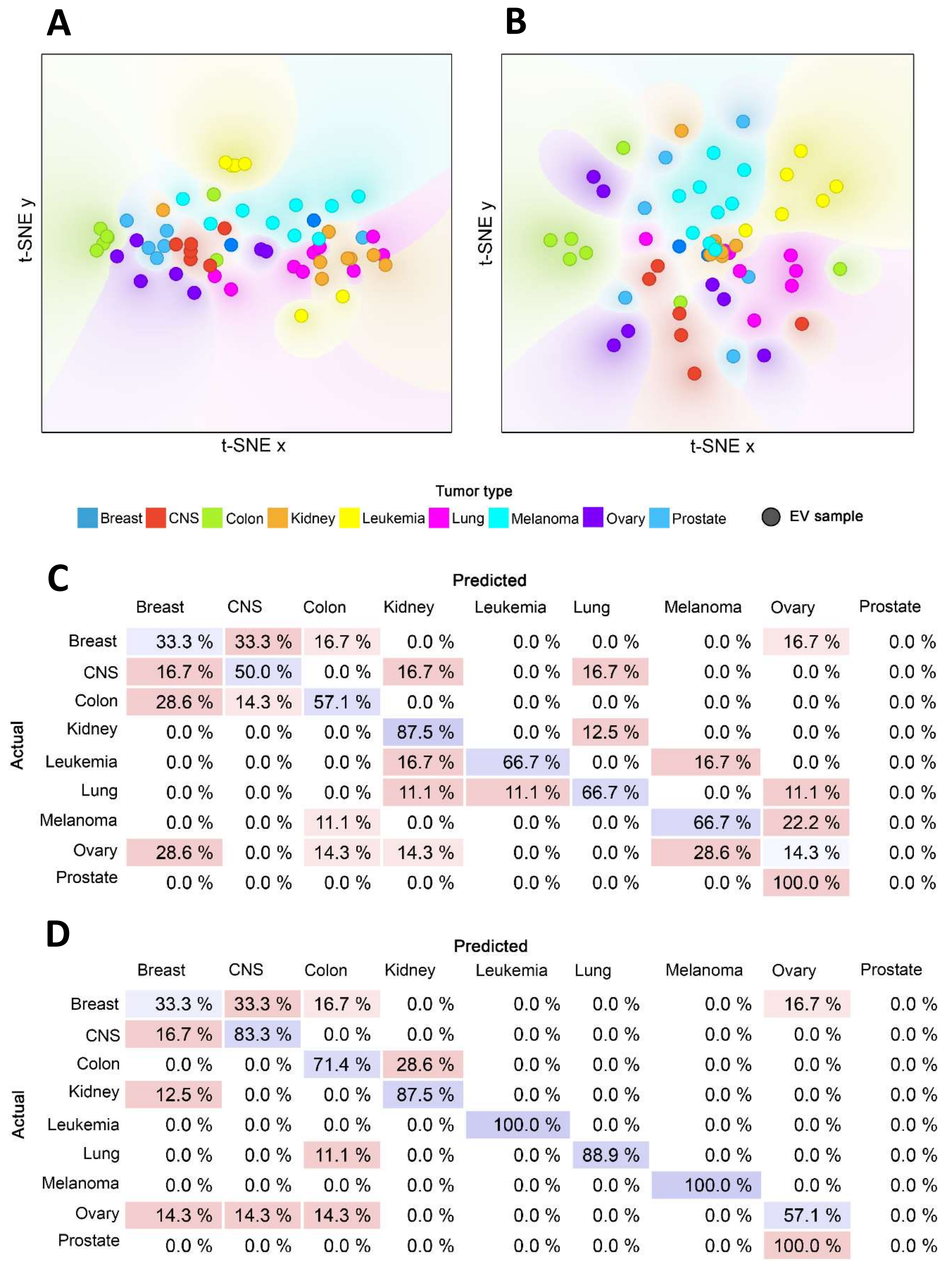
EV secretion, the mean and mode size of EVs, or cell size. No significant correlation was identified between any parameter and the average intensity of the core proteome. This suggests that the unique core proteome pattern is not caused by the difference in EV production rate and type of EVs or distinct cell size between the nine tumor types, but the different tissue origin.

When we expanded our focus to include the entire proteome, the distinction between tumor types had become even more define (**Figure 3B, D**). For CNS, colon, leukemia, lung, melanoma, and ovary tumors, classification accuracy improved significantly. The average classification accuracy for these types increased to 69.10%, which is 57.99% higher than what would be expected by chance.

4.1.3. The EV proteome could be used to form a discriminative protein panel

When exploring the discriminatory protein panel, we took care to avoid overestimation or overfitting of the method. To achieve this, we split the 60 cell lines into 50% to 50%. On one set, the Train set, we applied the LASSO algorithm.

Using the LASSO method, we were able to assign importance scores to each protein of the entire proteome based on their ability to differentiate the 9 tumor types in the Train set. The selection algorithm (with parameter C = 1) resulted in 172 proteins, which were further investigated for hierarchical clustering, classification purposes and Reactome pathway analysis (**Supplementary table 3**).



C

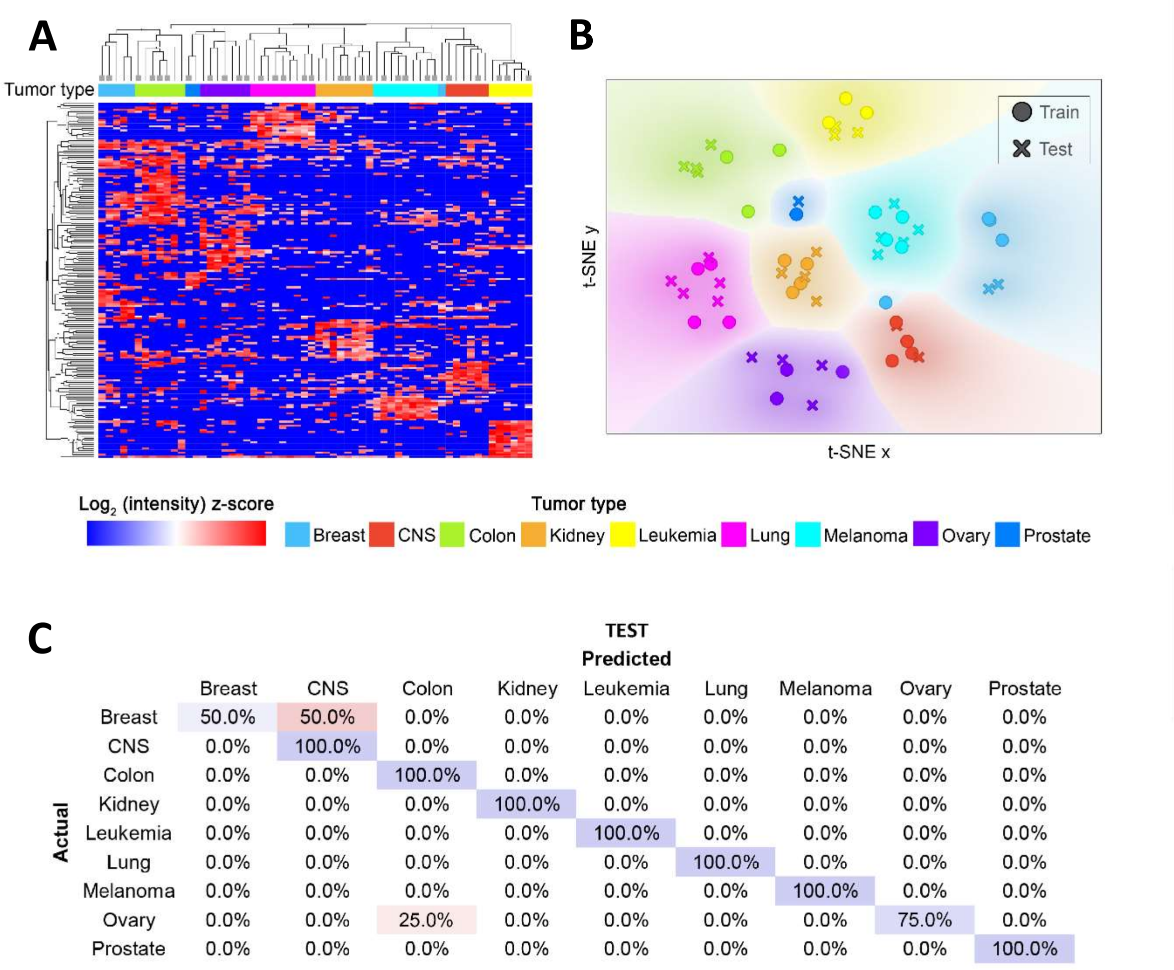
Actual \ Predicted	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	33.3 %	33.3 %	16.7 %	0.0 %	0.0 %	0.0 %	0.0 %	16.7 %	0.0 %
CNS	16.7 %	50.0 %	0.0 %	16.7 %	0.0 %	16.7 %	0.0 %	0.0 %	0.0 %
Colon	28.6 %	14.3 %	57.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Kidney	0.0 %	0.0 %	0.0 %	87.5 %	0.0 %	12.5 %	0.0 %	0.0 %	0.0 %
Leukemia	0.0 %	0.0 %	0.0 %	16.7 %	66.7 %	0.0 %	16.7 %	0.0 %	0.0 %
Lung	0.0 %	0.0 %	0.0 %	11.1 %	11.1 %	66.7 %	0.0 %	11.1 %	0.0 %
Melanoma	0.0 %	0.0 %	11.1 %	0.0 %	0.0 %	0.0 %	66.7 %	22.2 %	0.0 %
Ovary	28.6 %	0.0 %	14.3 %	14.3 %	0.0 %	0.0 %	28.6 %	14.3 %	0.0 %
Prostate	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %

D

Actual \ Predicted	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	33.3 %	33.3 %	16.7 %	0.0 %	0.0 %	0.0 %	0.0 %	16.7 %	0.0 %
CNS	16.7 %	83.3 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Colon	0.0 %	0.0 %	71.4 %	28.6 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Kidney	12.5 %	0.0 %	0.0 %	87.5 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Leukemia	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Lung	0.0 %	0.0 %	11.1 %	0.0 %	0.0 %	88.9 %	0.0 %	0.0 %	0.0 %
Melanoma	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %
Ovary	14.3 %	14.3 %	14.3 %	0.0 %	0.0 %	0.0 %	0.0 %	57.1 %	0.0 %
Prostate	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %

Figure 3. ***Classification efficiency based on the core and entire proteome.*** (**A**) t-SNE plot of the core proteome. (**B**) t-SNE plot of the entire proteome. The dots with different colors represent the 60 individual EV samples belonging to the nine tumor types. The color gradient in the plot indicates the dot density. (**C**) Confusion matrix of the classification results using the core proteome. (**D**) Confusion matrix of the classification results using the entire proteome. Each row of the matrices represents the instances in an actual class while each column represents the instances in a predicted class. Diagonally, the percentage of the correct classification is shown in blue. The percentage of errors is indicated in red. (CNS: central nervous system.)

In the hierarchical clustering, the Train and Test sets were analyzed together on the basis of 172 proteins. Hierarchical clustering using a heatmap revealed that the 172 proteins form a well-defined pattern, enabling the 60 EV samples to form nearly perfectly homogenous clusters, while the Train and Test sets elements are clustered together (**Figure 4A**)



TEST
Predicted

Actual	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CNS	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Colon	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Kidney	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Leukemia	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Lung	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Melanoma	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Ovary	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	0.0%	75.0%	0.0%
Prostate	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Figure 4. ***Classification efficiency for the selected proteins***. (**A**) Heatmap with hierarchical clustering. In the heatmap, the columns and rows represent the 60 EV samples belonging to the nine tumor types marked with different colors and the 172 proteins, respectively. Both the columns and rows are clustered. Dendrogram branches ending in a square indicate the elements to be included in the Train set. (**B**) t-SNE plot of the selected 172 proteins. The dots with different colors represent the 60 individual EV samples belonging to the nine tumor types. In the plot, the color gradient indicates the dot density. (**C**) Confusion matrix of the classification results using the selected proteins on the Test set. Each row of the matrices represents the instances in an actual class while each column represents the instances in a predicted class. Diagonally, the percentage of the correct classification is shown in blue. The percentage of errors is indicated in red. (CNS: central nervous system.)

This separation is also evident in the t-SNE plots, which depict the various tumor types as distinct groups (**Figure 4B**). Again, the elements of the Train and Test sets populated the same areas.

When the samples of the Test set were classified based on the 172 proteins, an average classification efficiency of 91.67% was achieved (**Figure 4C**).

4.1.4. Discriminative proteins might uncover tumor-specific pathways

After selecting the proteins, we hypothesized that – given the proteins' large intergroup differences – the biological signaling pathways they affect would also exhibit distinctive patterns. In order to place the 172 selected proteins in a biological context Reactome enrichment analysis was utilized. Only those pathways with $p <$ 0.05 were considered for hierarchical clustering and heatmap creation (**Figure 5**).

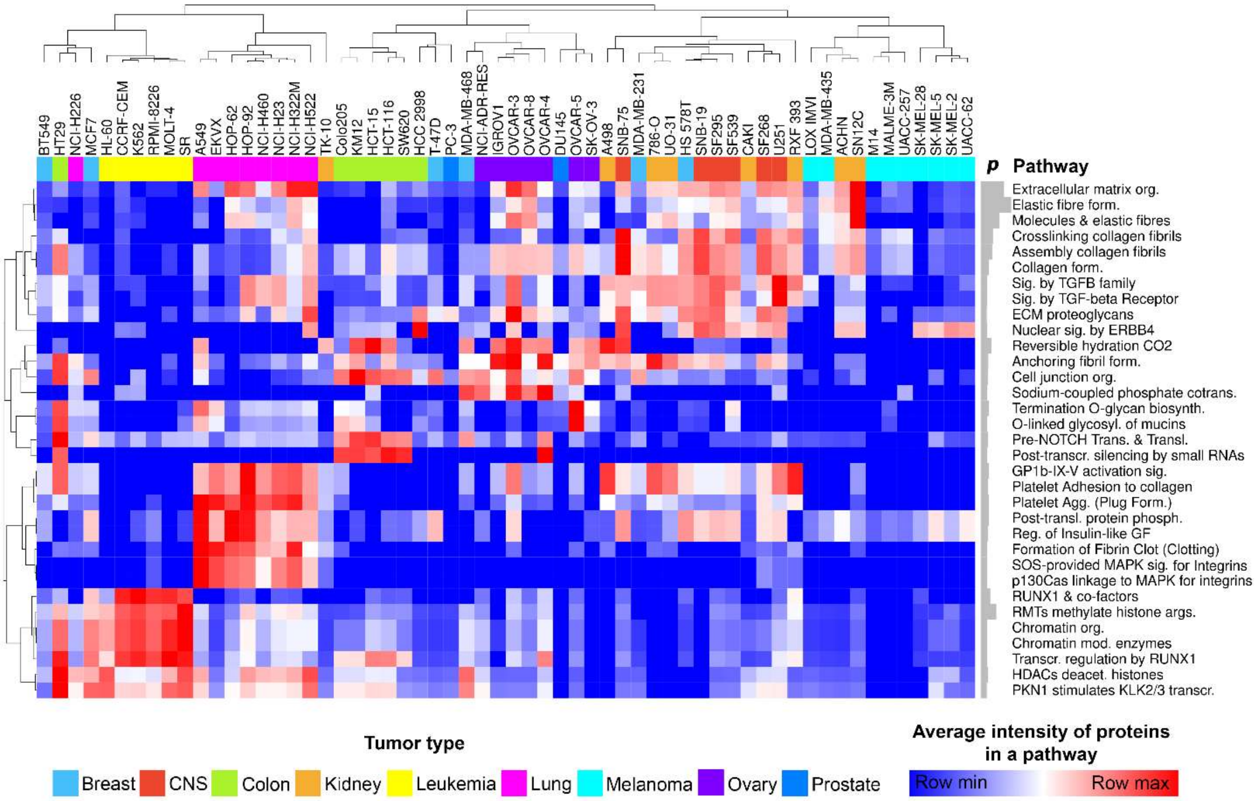


Figure 5. ***Biological signaling pathways affected by the 172 selected proteins of the discriminative protein panel.*** The columns marked with different colors represent the 60 EV samples, while the rows indicate the various signaling pathways. Both the 60 samples and pathways were clustered hierarchically. The heatmap values represent the average intensity of the proteins that are part of a given signal pathway. The gray barplots next to the names of the pathways indicate the $-\log_{10}$(*p* value). In all instances, $p < 0.05$. (agg.: aggregation; biosynth.: biosynthesis; cotrans.: cotransporters; deacet.: deacetylate; form.: formation; mod.: modifying; org.: organization; phosph.: phosphorylation; prots.: proteoglycans; sig.: signaling; trans.: transcription; transl.: translocation)

The selected 172 proteins are associated with extracellular matrix, nuclear processes, and cell division-related signaling pathways.

Although cancers of the breast and prostate lacked characteristic signaling pathways, the majority of the EV samples clustered according to their tumor type revealing a distinctive signaling pathway pattern.

The collagen matrix, TGF-β receptor, and ERB4 enzyme signaling pathways were identified as common characteristics for both kidney and central nervous system tumors, which clustered together.

Compared to other tumors, leukemia samples exhibit a predominance of nuclear processes associated with histone and chromatin modification.

In general, lung tumors were distinguished by platelet-associated biological processes and integrin-signaling pathways.

4.1.5. EVs carry information on invasion and proliferation capacity of donor cells

The NCI-60 cell line panel is diverse, comprising tumors of various tissue origins, each with differing capacities for invasion and rates of division. This prompted an inquiry into whether additional protein panels could be established to predict invasion and proliferation capacities. The question rose whether further protein panels predicting invasion and proliferation capacity could be determined.

To reveal such panels, we employed Multivariate linear regression combined with the LASSO selection method. The dataset was equally divided into two halves, with the Train set used for initial analysis. In this set, the LASSO method identified proteins that could potentially predict invasion capacity and proliferation rates. These findings were then tested for validity on the Test set.

The process resulted in the identification of 20 proteins related to invasion capacity and 15 to proliferation capacity in the Train set, forming separate panels for invasion and proliferation. The predictive value of these panels was then assessed on the Test set using Multivariate linear regression.

This analysis yielded significant results for both panels ($p < 0.0001$), with remarkably high coefficients of determination: $R^2 = 0.68$ for invasion and $R^2 = 0.62$ for proliferation (**Figure 6**).

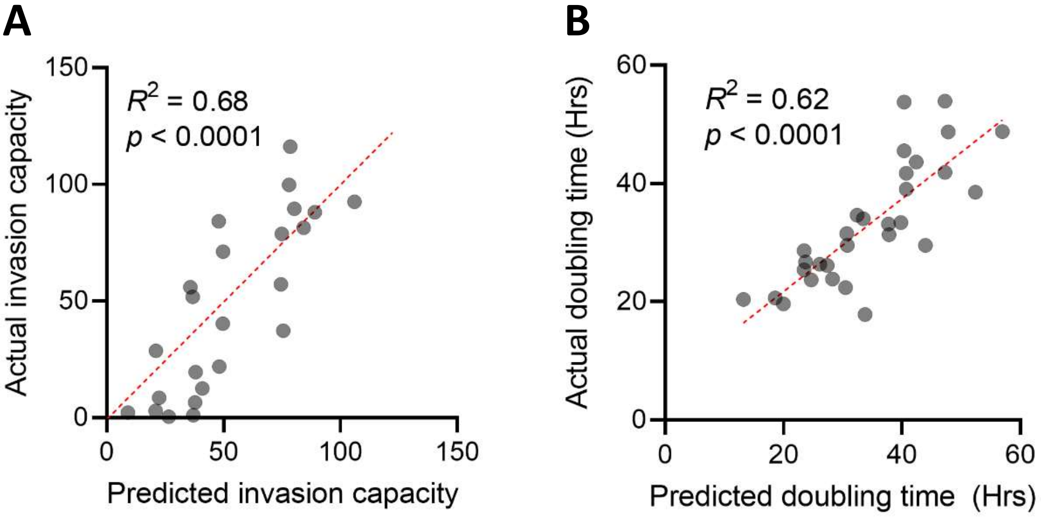


Figure 6. ***Results of the Multivariate linear regression.*** (**A**) Multivariate linear regression of invasion capacity. The invasion capacity predicted by the invasion panel for each sample in the Test set is plotted on the x-axis, while the actual invasion capacity is plotted on the y-axis. (**B**) Multivariate linear regression of proliferation capacity. The doubling time predicted by the invasion panel for each sample in the Test set is plotted on the x-axis, while the actual doubling time is plotted on the y-axis. (R^2: coefficient of determination; *p*: *p* value.)

Post-validation on the Test set, which confirmed the predictive accuracy of these proteins, both the 20-protein and 15-protein panels underwent hierarchical clustering. This resulted in two distinct clusters for each panel (**Figure 7**): one cluster seemingly negatively correlated and the other positively correlated with either invasion or proliferation capacity.

Of the 20-member invasion panel, 8 proteins (CAV2, DNAJB4, THY1, OXTR, VCAN, COL11A1, EDIL3, CRYAB) positively predicted the invasion capacity of the cell lines. Based on Reactome pathway analysis, these proteins were significantly associated with signaling pathways that upregulate tumor cell maintenance, invasion and binding to the extracellular matrix. Similarly, the enrichment analysis of the remaining 12 proteins (RGS19, SLCA43A3, MYO18A, HIST1H3A, PSME3, SYNGR2, PPP1CC, FIS1, PARP1, VPS11, TGFBRAP1) that negatively predict invasion capacity was consistent with the regression results: these proteins play a role in pathways that negatively regulate the invasion (**Figure 7A**).

The 8 proteins (FRYL, HDAC1, ANAPC1, THOC2, WDR12, DSG2, SLC38A5, SLC38A1) that positively influence proliferative capacity were associated with processes linked to cell cycle. While 7 proteins (TBC1D2, FKPB2, ECH1, GLUD1, ACOT13, ERGIC1, PDZK1IP1) negatively associated with proliferation are linked to metabolic pathways (**Figure 7B**).

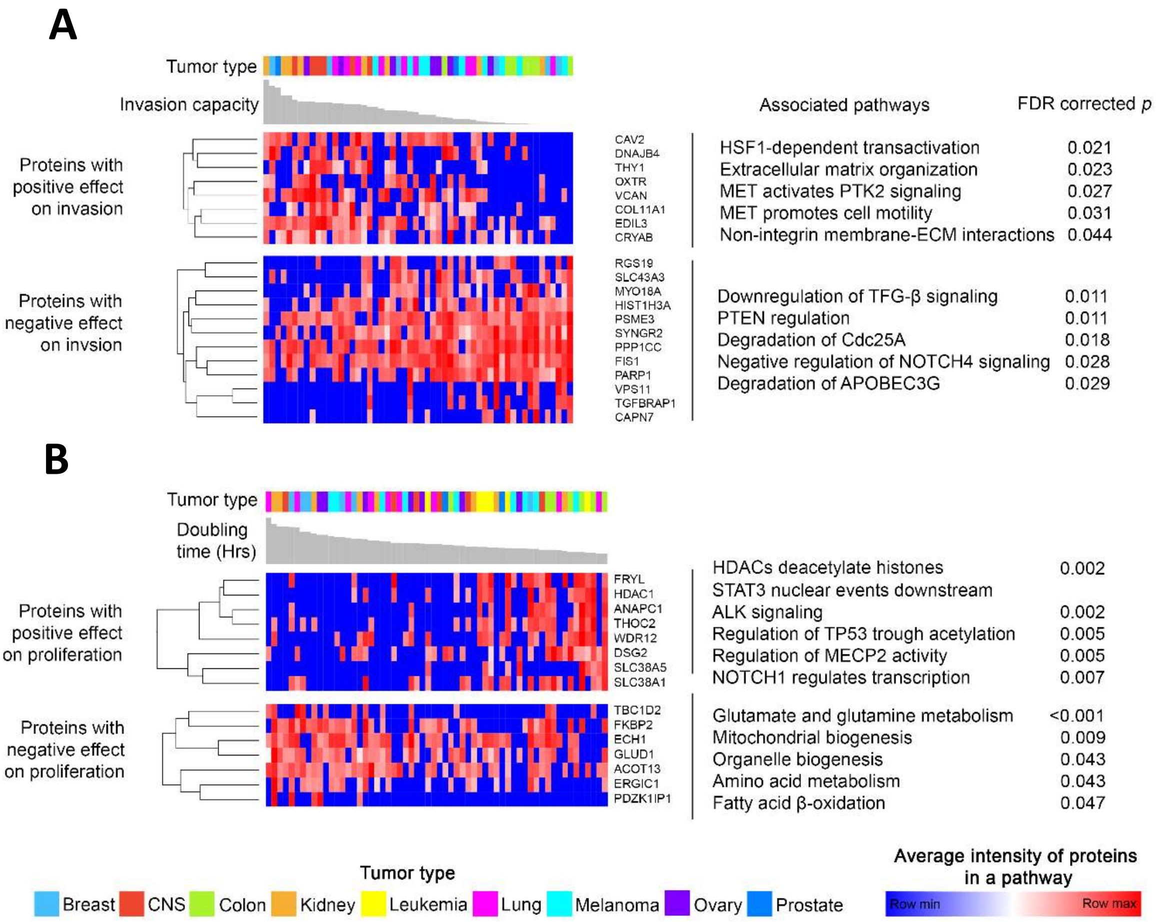


Figure 7. ***Predictive proteins for invasion and proliferation capacity***. (**A**) Predictive protein panel for invasion capacity (invasion panel). The columns marked with different colors and the gray barplots indicate the 54 EV samples with the invasion capacity measured for the cell line of origin (leukemia not included). The rows indicate the proteins, which were clustered hierarchically. Two defined clusters were separated from each other. (**B**) Predictive protein panel for proliferation capacity (proliferation panel). The columns marked with different colors and the gray barplots indicate the 60 EV samples with the doubling time (in hours) measured for the cell line of origin. The rows indicate the proteins, which were clustered hierarchically. Two defined clusters were separated from each other. It should be noted that higher doubling time means lower proliferation capacity as it indicates more time for cell division. (FDR: False discovery rate; p: *p* value)

We further attempted to gain more support for our invasion and proliferation capacity prediction panels by examining their impact on patients' survival time.

The HPA was considered an appropriate database for this purpose, as it contains survival times for a large number of cancer patients for all 9 cancer types and is easily accessible. However, we had to take into account the limitation that HPA contains tissue RNA expression data and not EV proteomic data.

Accordingly, before utilizing the HPA database, we had to assess the similarity of EV protein and cellular RNA patterns to be permitted to investigate the effect of *in situ* RNA tissue expression of panel members on survival time.

First, we examined how the EV protein panels (invasion and proliferation) and the cellular RNAs correlate with each other (**Figure 8**). Based on the results, the RNA and protein patterns of the invasion panel showed a moderately strong concordance ($RV = 0.51$, $p = 0.020$). While a weaker but still significant relationship was observed when comparing the RNA and protein matrices of the proliferation panel ($RV = 0.39$, $p = 0.048$). Notably, we observed stronger pairwise correlations between protein and RNA content for the promoting members of both panels.

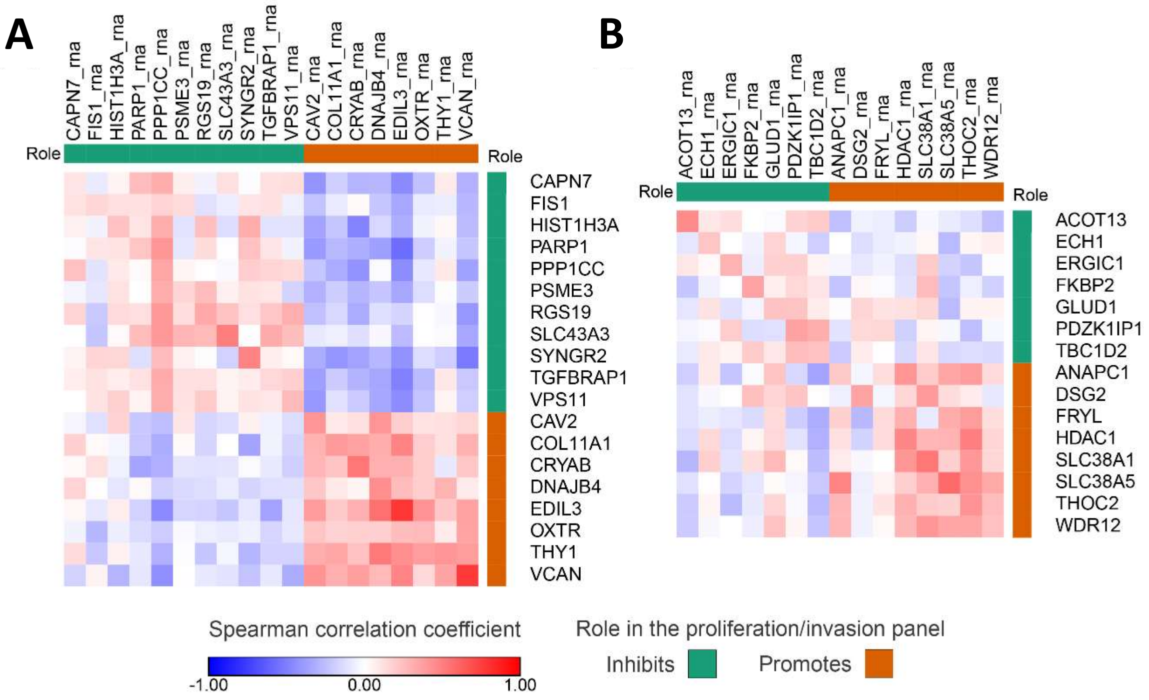


Figure 8. ***Correlation of EV protein and cellular RNA content***. The heatmaps show the correlation between cellular RNAs and EV proteins of invasion (**A**) and proliferation (**B**) panel members. Columns represent the cellular RNA, rows represent the EV proteins. (See ***Abbreviations*** for an explanation of molecule names.)

After assessing the relationship between EV protein and cellular RNA pattern, we attempted to use the cellular RNA to estimate the invasion and proliferation capacity of cells using the panel members.

Based on the cellular RNA, invasion capacity could be estimated at $R^2 = 0.77$ ($p < 0.0001$) and proliferation capacity at $R^2 = 0.32$ ($p = 0.037$).

The in vitro data suggested that the EV proteomic and cellular RNA patterns are in concordance and that the cellular RNA content is also related to invasion and proliferation capacity in a similar way as the EV proteome. This prompted us to investigate the impact of in vivo RNA tissue expression of panel members on patient survival.

Using the HPA database, we collected clinical data on the tissue expression of our panel members in the 9 tumor types from 4,665 patients, then examined the relationship between tissue expression and 5-year survival rate.

In the HPA database, tissue expression was found for 19 of the 20 proteins of the invasion panel (**Supplementary table 4**).

According to the HPA, high expression of CAV2, COL11A1, DNAJB4, THY1 and VCAN decreased the 5-year survival for breast, CNS, colon, kidney, lung and ovarian tumors (**Figure 9A**). These findings are in line with our results, as these proteins were found to be positively associated with invasion capacity according to multivariate linear regression analysis.

The CRYAB protein was found to be controversial, as our results showed a positive association with invasion, but in HPA, high tissue expression was associated with a better prognosis in CNS tumors. Nevertheless, in colon tumors, high expression was a negative prognostic marker.

The case is similar for EDIL3, which is positively associated with invasion capacity according to Multivariate linear regression analysis, but based on the HPA, higher tissue ex-pression is associated with better 5-year survival in colon tumors. However, it still was a significantly worse prognostic marker in breast, kidney, and melanoma patients.

Overall, the effects on survival found in the HPA database and the effect of the proteins on invasion capacity as determined in our study were consistent in 90% of the cases.

Based on multivariate linear regression, 12 of 20 proteins in our study were found to be negatively correlated with invasion capacity (**Figure 9B**). Comparing this

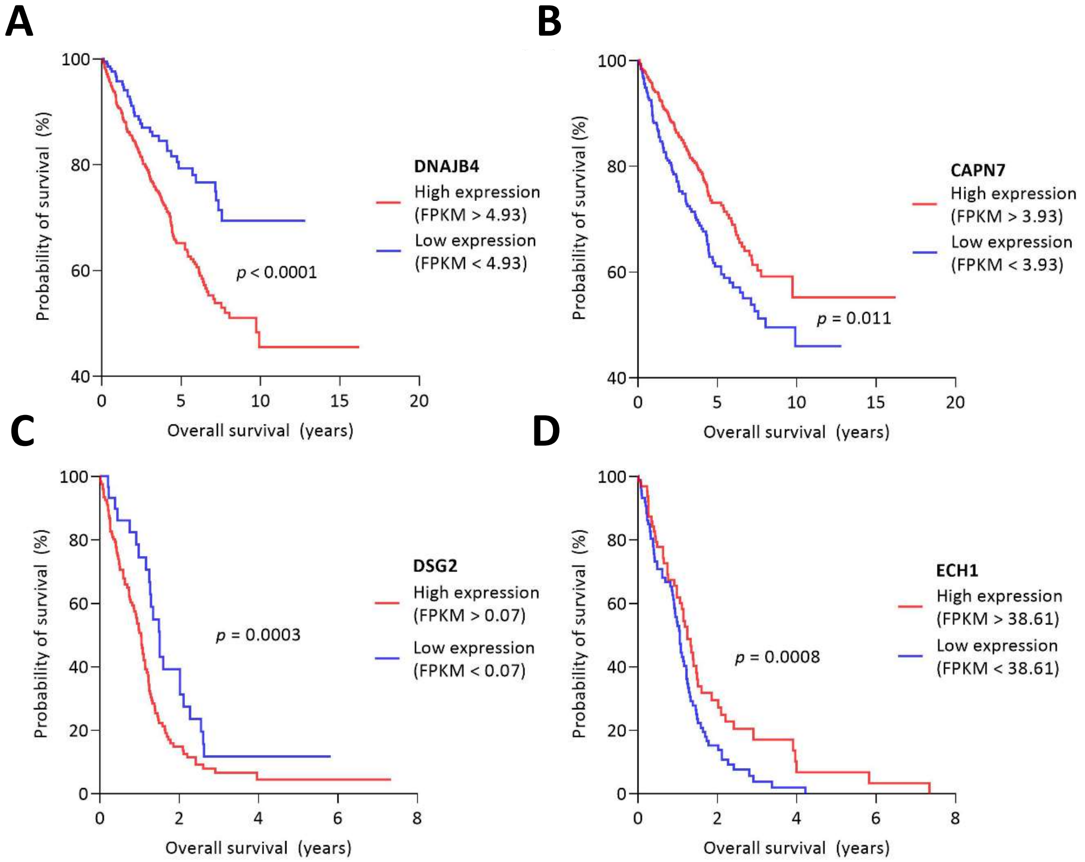


Figure 9. ***Survival functions for different expression levels of DNAJB4, CAPN7, DSG2, ECH1***. The figure shows 4 exemplary proteins selected from the members of the invasion and proliferation panel and their impact on patients' survival. (**A**) DNAJB4, which we found to be positively associated with invasion and which the Human Protein Atlas (HPA) suggests that its high expression is associated with a worse prognosis in kidney tumors ($n = 877$). (**B**) CAPN7 protein, which in our study is negatively associated with invasion and which the HPA suggests may be associated with a favorable prognosis in kidney tumors. (**C**) DSG2 protein which in our study positively predicted the proliferation capacity is a negative prognostic factor in CNS tumors, based on HPA. (**D**) Based on our results, ECH1 protein negatively predicted the proliferation capacity, and it is a favorable prognostic marker for CNS tumors. (FPKM: Fragment per kilobase per million).

finding to the HPA database, we found more inconsistencies: according to the HPA, the 12 proteins are favored prognostic markers for 5-year survival in most cases (73.18%) but in 26.82%, the proteins have an adverse effect on survival than the expected. For example, HIST1H3A showed a negative association with invasiveness in our study, but its high expression negatively affected the survival rate of CNS tumor patients according to the HPA database (**Supplementary table 4**).

Tissue expression was found for all the 15 proteins of the proliferation panel (**Supplementary table 5**). The high expression of 8 out15 proteins, which positively predict the proliferation capacity, significantly reduces the 5-year survival in 72.41%

of cases (**Figure 9C**). The proliferation panel contains 7 out 15 proteins which were found to negatively predict the proliferation capacity. According to HPA, high tissue expression of these 7 proteins significantly increased the 5-year survival in 64.71% of cases (**Figure 9D**).

Taken as a whole, the EV proteome and in vitro cellular RNA pattern of the panel members showed concordance, and the effect of in vivo tissue RNA expression of the panel members on patient survival is consistent with the results of our linear regression model. The finding potentially suggests the involvement of invasion and proliferation panels in the tumorous processes.

It is noteworthy that the inconsistency with HPA appears for those variables where the in vitro EV proteome and cellular RNA pattern did not show a strong correlation (invasion capacity inhibitory members) (**Figure 8**), or cellular RNA did not prove to be a sufficient predictor (overall the proliferation panel).

4.2. Finding of the clinical study on serum-derived extracellular vesicles

4.2.1 Particles isolated from serum show sEV properties

Particles were isolated by differential centrifugation from 138 serum samples of patients with GBM, BM, M and CTRL. Isolated particles were characterized by TEM and NTA, as well as by examining characteristic sEV markers (Alix, CD81 and calnexin) by WB (**Figure 10**). Average concentration, mean and mode diameter of the particles were measured as 7.41 × 10^{10} particles/mL, 111.20 nm and 83.32 nm, respectively. Alix, CD81 positivity and calnexin negativity was determined. The results of the characterisation showed that the isolate was enriched in particles that fall into the sEV category. No statistically significant differences were identified among the patient groups in any of the parameters of the isolated particles.

In our clinical study, considering the presence of non-vesicular elements in the samples after isolation, such as protein aggregates, apolipoproteins, abundant serum proteins, our samples are not considered as pure sEV isolates, but are referred to as sEV-enriched isolates.

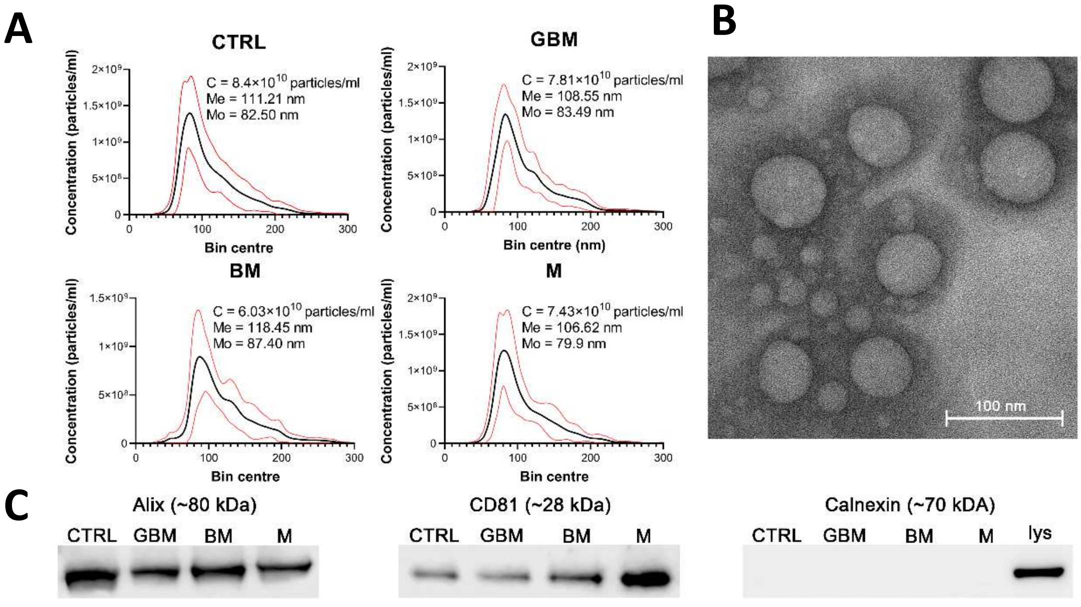


Figure 10. ***Characterization of the particles.*** The figure represents the results of the particle characterization: size distribution of the sEV samples isolated from the four patient groups (black and red lines represent the mean and the standard deviation of the concentration, respectively) (**A**), a representative TEM image of the sEVs (**B**), and the Western blot analysis of the sEV markers (**C**). (CTRL: control, GBM: glioblastoma multiforme, BM: brain metastasis, M: meningioma, C: particle concentration, lys: cell lysate, Me: mean diameter size, Mo: mode diameter size.)

4.2.2. Patient groups can be distinguished using the PCA–SVM algorithm with high classification efficiency

Raman spectroscopic analyses of the isolated 138 samples yielded 5 spectra per sample. The spectral range between 799.5 cm^{-1} and 3100.5 cm^{-1} was investigated. After SNV normalization and PCA transformation, the classification of samples was performed using the SVM algorithm. Classification efficiency was evaluated by classification accuracy (CA), sensitivity, specificity and the area under the curve (AUC) value derived from the ROC analysis. Relevant spectral differences were revealed by PCA. **Figure 11** shows the flowchart of Raman spectroscopy data processing.

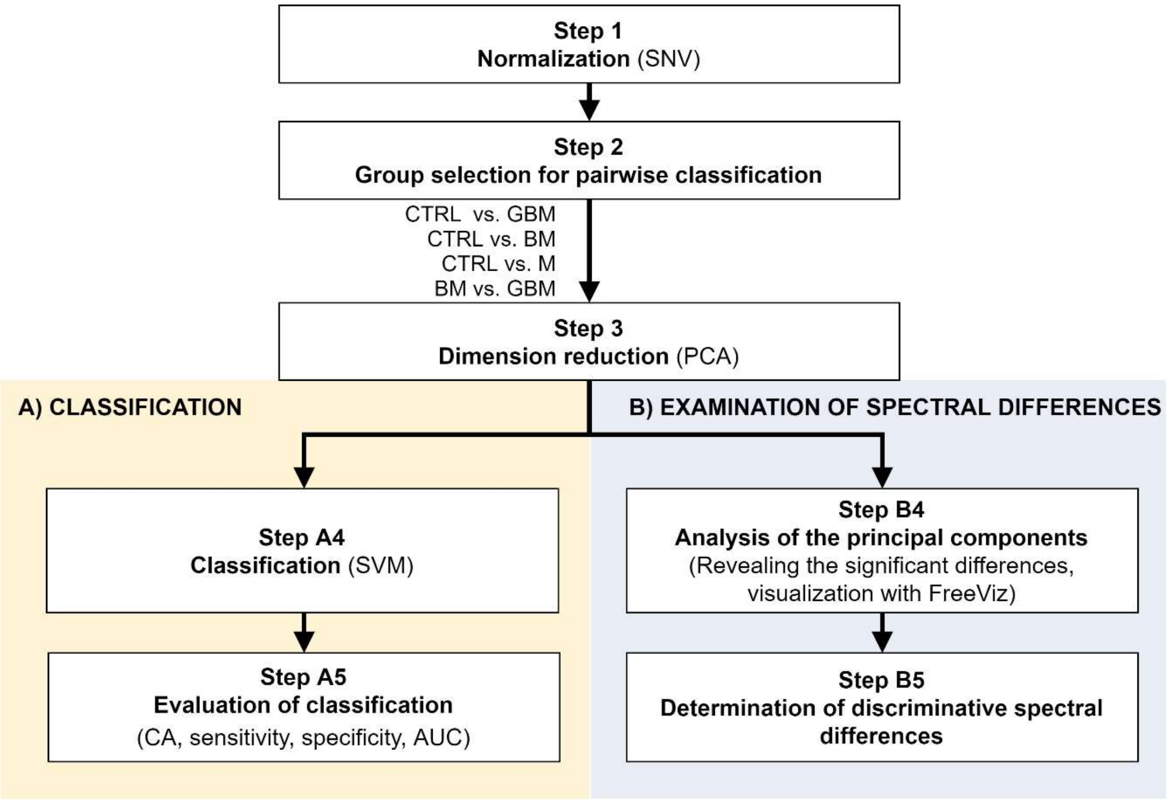

Step 1
Normalization (SNV)
Step 2
Group selection for pairwise classification
CTRL vs. GBM
CTRL vs. BM
CTRL vs. M
BM vs. GBM
Step 3
Dimension reduction (PCA)
A) CLASSIFICATION
B) EXAMINATION OF SPECTRAL DIFFERENCES
Step A4
Classification (SVM)
Step A5
Evaluation of classification
(CA, sensitivity, specificity, AUC)
Step B4
Analysis of the principal components
(Revealing the significant differences,
visualization with FreeViz)
Step B5
Determination of discriminative spectral
differences

Figure 11. ***Workflow of Raman spectroscopic data processing.*** The figure shows the analysis step by step. After Step 3, the workflow separates (parts **A** and **B**) according to the purpose of the analysis. (AUC: area under the curve; CA: classification accuracy, SNV: standard normal variate, SVM: support-vector machine, PCA: principal component analysis).

After averaging the spectra, row normalization was performed using the SNV method (Step 1) (**Figure 12**).

Following SNV-normalization, the spectra for the samples of the four patient groups were compared pairwise (each patient group was compared to the control, and BM vs. GBM was compared) for two purposes: first, to develop and test a classification algorithm, and second, to identify relevant spectral differences. PCA applied on the pairwise comparisons reduced multivariate data dimensions by transforming the original variables (wavenumbers) into a smaller number of new variables, i.e., principal components (PCs) (Step 3).

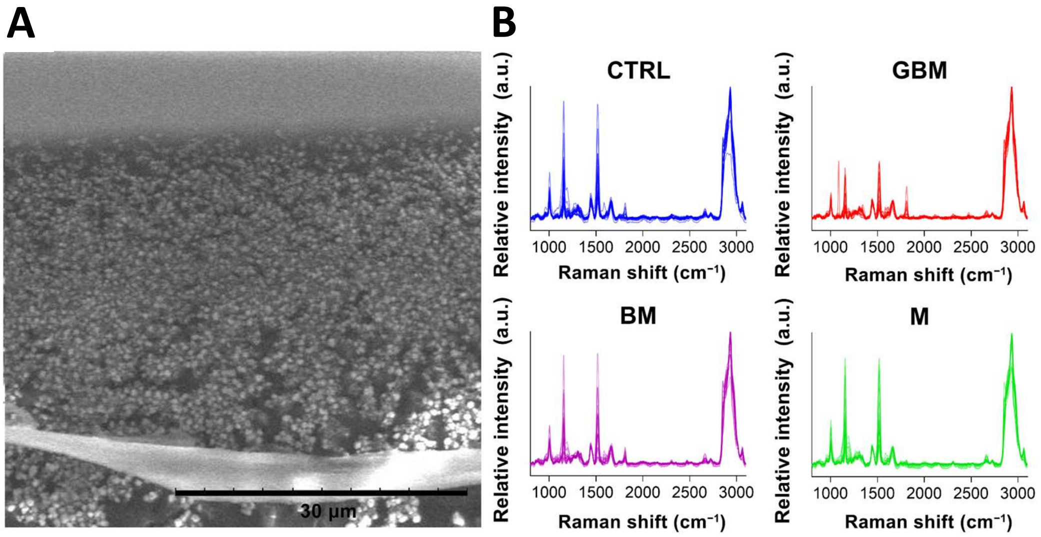


Figure 12. ***EVs under Raman spectrometer.*** Particles on a calcium fluoride substrate (scanning electron microscope image) (A). Averaged and SNV-normalized spectra of the four patient groups (B). (a.u.: arbitrary unit, BM: brain metastasis of non-small cell lung cancer, CTRL: control, lumbar disc hernia, GBM: glioblastoma multiforme, M: meningioma).

Pairwise comparisons were conducted using the linear SVM (Step A4) algorithm, yielding classification models for each paired group. To make predictions for the test samples, a minimum threshold for the group-membership score was determined. Test samples with scores above this threshold were classified into the target group of interest. The optimal score thresholds were automatically set to correspond to the highest classification accuracy (CA, the ratio of correctly classified samples per all samples).

CA was 85.6% for CTRL vs. GBM, 91.4% for CTRL vs. BM, 82.9% for CTRL vs. M and 92.5% for BM vs. GBM. The best classification performance was achieved when a certain number of PCs were included in the models: 30 PCs for CTRL vs. GBM, 38 PCs for CTRL vs. BM, 27 PCs for CTRL vs. M, and 26 PCs for BM vs. GBM (**Supplementary figure 1**).

Sensitivity and specificity were evaluated as further metrics of classification performance. ROC analyses of the pairwise classification models yielded four graphs showing the automatically set optimal thresholds (having the highest CA value), with related sensitivity, specificity and AUC values, as well as *p* value (**Figure 13**).

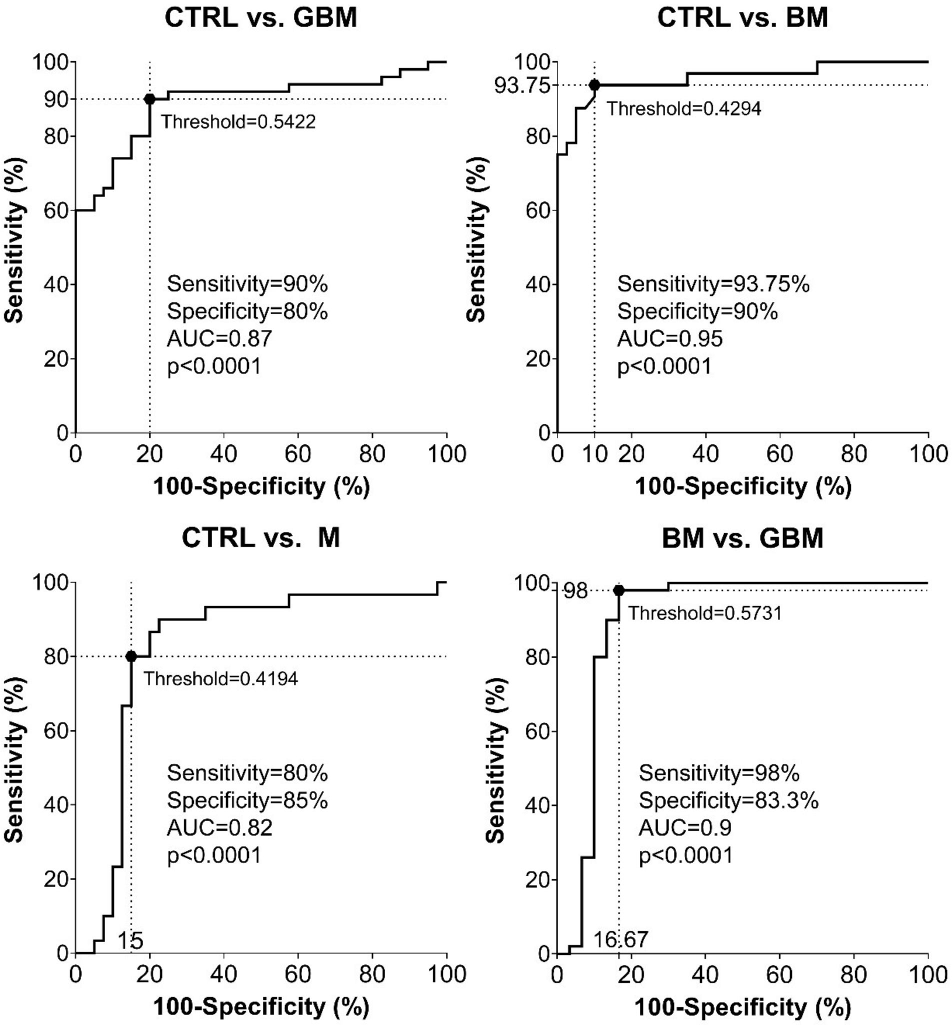


Figure 13. ***Receiver Operating Characteristic (ROC) curves for the classification models.*** Intersecting black dotted lines show sensitivity, specificity and corresponding threshold values of the group-membership score, with black filled circles at their intersections. (AUC: area under the curve, BM: brain metastasis of non-small cell lung cancer, CTRL: control, lumbar disc hernia, GBM: glioblastoma multiforme, M: meningioma, p: *p* value from ROC analysis).

As shown in the graphs in **Figure 13** using the optimal thresholds, the classification models were able to distinguish GBM, BM and M patients from CTRL patients with a sensitivity and specificity of 90% and 80%, 93.75% and 90%, 80% and 85%, respectively (Step A5). Using the classification model, the two malignancies, BM and GBM, could be distinguished from each other with a sensitivity of 98% and a specificity of 83.3%. In the same order of pairwise comparisons (GBM, BM and M patients vs. CTRL, and BM vs. GBM), the AUC values were 0.87, 0.95, 0.82 and 0.9, respectively ($p < 0.0001$ in all cases).

4.2.3. Analysis of the PCs revealed discriminative spectral differences

Next, differences in the molecular content of serum-derived sEV-enriched isolates from each group were investigated to reveal the spectral differences relevant with regard to the classification. SNV-normalized spectra and the PCs obtained from PCA were analyzed using the FreeViz method, to reveal and visualize relevant spectral differences (**Figure 14**).

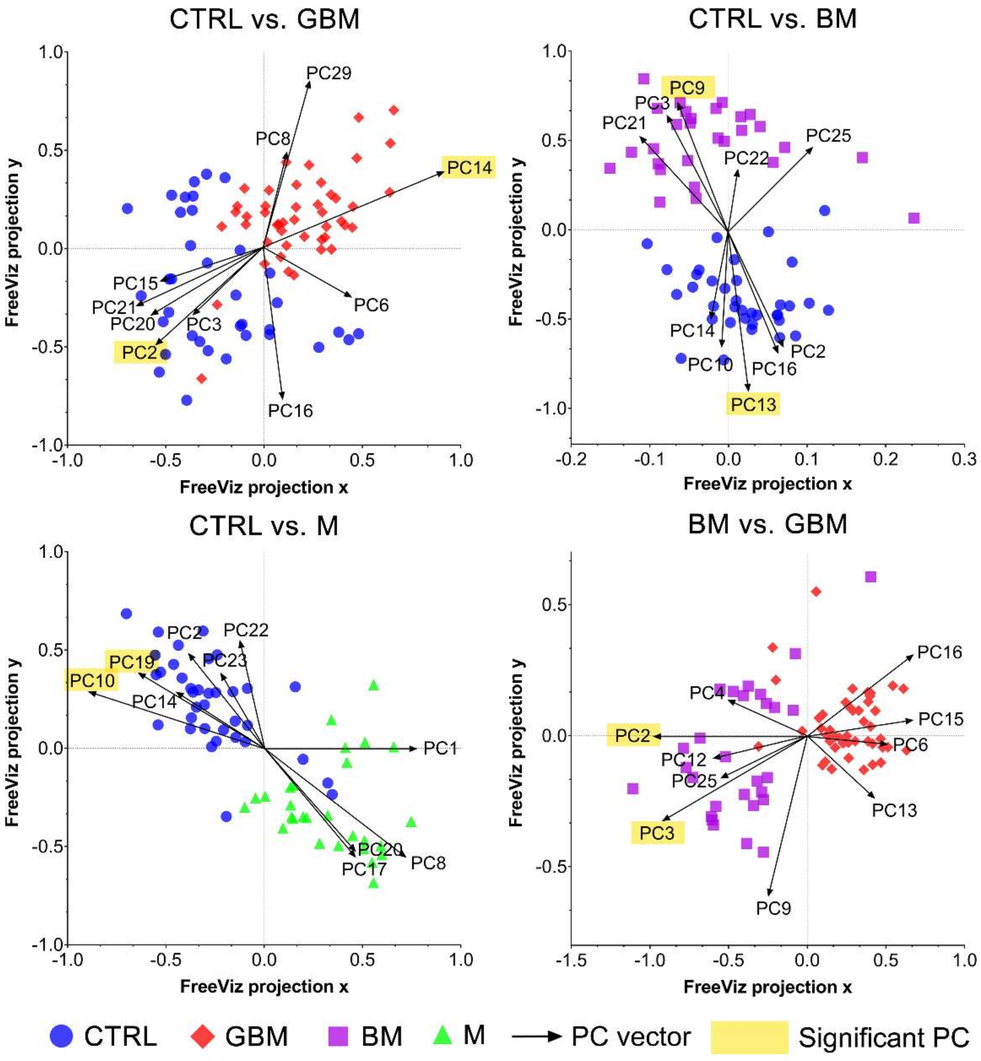


Figure 14. ***FreeViz projections of pairwise comparisons.*** Analysis of the PCA-transformed data using the FreeViz method yielded four graphs. Different dots and colors represent the patient groups and healthy controls. Black vectors represent the PCs. In each graph, only the 10 most relevant PC vectors were plotted. For each comparison, PCs marked with a yellow background indicate the 2 most significant PCs. (BM: brain metastasis of non-small cell lung cancer, CTRL: control, lumbar disc hernia, GBM: glioblastoma multiforme, M: meningioma, PC: principal component)

The FreeViz method (Step B4) displayed the optimized projections of the multivariate data sets in a 2-dimensional scatterplot (**Figure 14**). Based on the length and direction of PC vectors, two PCs that were revealed to play the most important role in distinguishing each paired group (marked with a yellow background in **Figure 14**) were further assessed to determine discriminative spectral signatures.

Based on the results of the FreeViz method and *p* values from Welch's *t*-test, PC14 and PC2, PC9 and PC13, P10 and PC19, and PC2 and PC3 explained most of the discriminative differences in the CTRL vs. GBM, CTRL vs. BM, CTRL vs. M and BM vs. GBM comparisons, respectively ($p < 0.05$ in all cases).

Evaluating the selected PCs, we attempted to find the chemical bonds and functional groups corresponding to the spectral differences found to have an important role in distinguishing the compared groups (Step B5) (**Figure 15**).

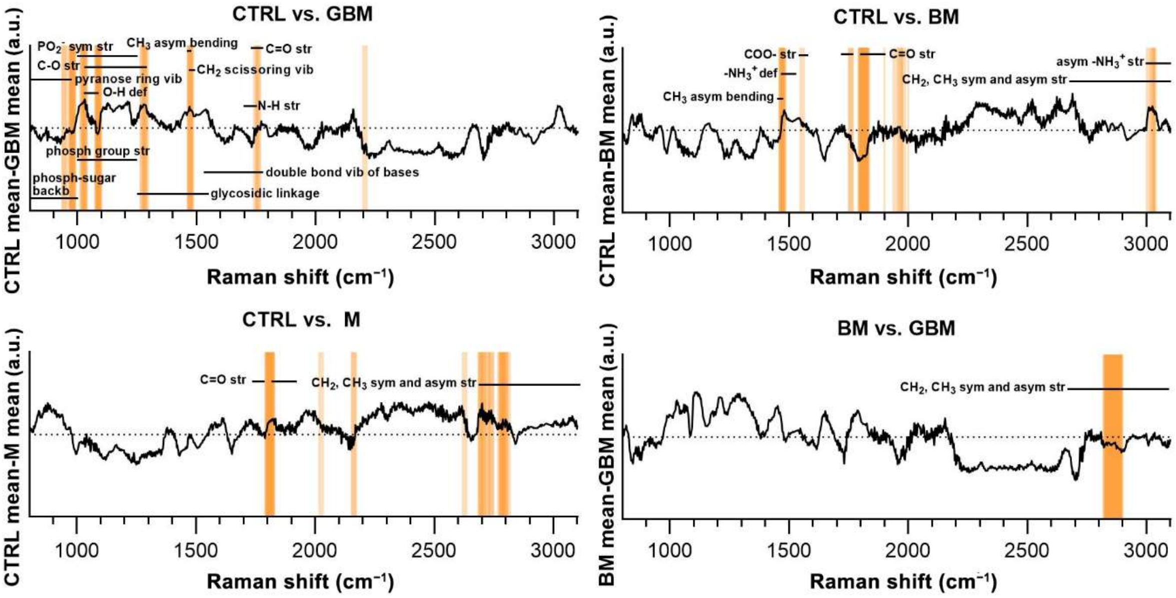


Figure 15. ***Subtraction spectra for the pairwise comparisons.*** Subtraction spectra were produced by subtracting the mean signal intensities for the groups compared. Spectral regions having a higher-than-average contribution to significant PCs were marked with orange bars. The more saturated a bar is, the more that region is represented on the selected PCs. The dotted horizontal line represents zero difference at y = 0. (asym: asymmetric, backb: backbone, BM: brain metastasis of non-small cell lung cancer, CTRL: control, lumbar disc hernia, def: deformation, GBM: glioblastoma multiforme, M: meningioma, phosph: phosphate, str: stretching, sym: symmetric, vib: vibration).

Regarding the CTRL vs. GBM comparison, most of the discriminative spectral differences were characteristic for carbohydrates, such as bands associated with a pyranose ring (800–975 cm^{-1}), O-H deformation vibrations (1030–1080 cm^{-1}) and C-O stretching vibrations (1030–1290 cm^{-1}). These bands largely overlap with the region's characteristic for nucleic acids, including the bands associated with the vibrations of the phosphate-sugar backbone (800–1000 cm^{-1}), symmetric and asymmetric phosphate group stretching vibrations (1000–1250 cm^{-1}), glycosidic bond vibrations (1250–1550 cm^{-1}), and in-plane double bond vibrations of bases (1530–1780 cm^{-1}) (**Figure 15**).

Regarding the CTRL vs. BM comparison, the wavenumbers found to have an important role in distinguishing the BM group from the control mainly correlated with lipids (CH_3 asymmetrical bending (1470–1490 cm^{-1}), CH_2 and CH_3 symmetrical and asymmetrical stretching vibrations (2700–3100 cm^{-1})) and amino acids ($–NH_3^+$ deformation band (1485–1150 cm^{-1}), $–NH_3^+$ asymmetrical stretching (3000–3100 cm^{-1}), carboxylate ion stretching (1560–1600 cm^{-1}) and C=O stretching vibrations of the carboxyl group (1700–1755 cm^{-1})). Regarding the CTRL vs. M and BM vs. GBM comparisons, the wavenumbers highly correlated with vibrations originating from acyl chains of lipids, such as CH_3 and CH_2 symmetric and asymmetric stretching vibrations (2700–3100 cm^{-1}).

5. DISCUSSION

In the following, the ***Discussion*** synthesizes the findings of the underlying meta-analysis and clinical studies and draws conclusions regarding the benefits of machine learning in both research, the potential role of EVs in tumor diagnostics, differential diagnosis, better understanding of cancer characteristics.

5.1. Machine learning methods tailored to prediction and classification problems

A wide range of machine learning methods can be applied to explore the relationships in biological "big data". In our research we used both classification algorithms and methods to select the relevant variables.

The application of linear and logistic regression coupled with LASSO, and the SVM algorithm, proved to be an optimal choice from both theoretical and practical perspectives.

LASSO has shown great promise in the analysis of proteomic data sets[83,84]. One of the key advantages of LASSO is its capability for variable selection: in regression analyses, it effectively shrinks the coefficients of less important variables to zero, thereby eliminating them from the model. This feature is especially valuable in the field of proteomics, where data sets often comprise a large number of variables, not all of which are pertinent. By reducing the number of variables, LASSO facilitates the creation of simpler, more interpretable, and more generalizable models[84].

The benefits of LASSO have been demonstrated in various medical research. By employing LASSO, researchers can pinpoint critical genes linked to immune cell infiltration in pediatric septic shock[85]; it also supports the creation of molecular panels for predicting outcomes in triple-negative breast cancer[86], or diagnose colon cancer[87] and bladder cancer[88].

In our clinical study, we utilized SVM for classification problems, particularly due to its reliable theoretical foundation[89]. SVM have emerged as a prominent tool for classification tasks involving spectral data, primarily due to their robust handling of outliers and applicability in scenarios where variables outnumber samples—a common situation in spectral analysis. The efficacy of SVM in spectral data analysis is underscored by numerous studies. For instance, Zhang et al. achieved over 90% efficiency in distinguishing breast cancer subtypes[90], while Huang et al. similarly excelled in classifying esophageal squamous carcinoma[91]. Beyond distinguishing healthy from tumor samples, SVM has also been utilized to classify different types of tumors with an accuracy of 80% based on Raman spectra[92].

Our study, incorporating LASSO and SVM revealed high tumor specificity of protein pattern carried by EVs, enabled predictions about the invasion capacity and proliferation time of donor cells, provided insights into tumor characteristic pathways, and efficiently classified CNS tumor patients based on Raman spectra of sEV-enriched isolates from serum samples.

5.2. Raman spectroscopy of EVs as a promising diagnostic approach

The diagnostic potential of EVs in identifying tumors has been highlighted by our recent clinical research, which specifically focuses on central nervous system (CNS) tumors. In this study, we isolated and examined serum-derived EVs from patients diagnosed with glioblastoma multiforme (GBM), meningioma (M), and brain metastases originating from NSCLC (BM). We also included patients with lumbar disc hernia as a control group (CTRL).

While EVs have been recognized as promising biomarkers for a variety of malignancies—including cancers of the head and neck, lung, prostate, skin, breast, and colorectum—their diagnostic efficacy for CNS tumors remains underexplored, lags behind other cancer types[33,34,36–38,43,65–67,93–95]. Prior studies have primarily focused on the nucleic acid, protein, metabolite, or lipid contents of circulating EVs from CNS tumor patients, often targeting a limited number of biomarkers. Unfortunately, these stand-alone biomarkers from different omics or small biomarker sets have not demonstrated sufficient diagnostic or prognostic value to be of clinical use. Furthermore, the validity of these biomarkers has yet to be confirmed using blinded clinical samples [33,34,36–38,43,65–67,93–95].

In pursuit of a more comprehensive approach, our research leverages the entire molecular profiling of sEV-enriched isolates with Raman spectroscopy. The effectiveness of our approach is well demonstrated by the machine learning-based classification models we have developed. These models, according to ROC analysis standards, have been rated as "excellent" and "outstanding" in distinguishing between CTRL and each of GBM, M, and BM[96].

The results obtained in the GBM vs. CTRL comparisons are considered to be more noteworthy as GBM is known to have a high variability among patients, which makes it difficult to identify a diagnostic biomarker that can be used generally in the population. Yet, the ability of our model to discriminate CTRL from GBM based on Raman spectra is outstanding (sensitivity = 90%, specificity = 80%), which puts it at the forefront of diagnostic efficiency reported in the last five years[97–101].

Similar to GBM, research on liquid biopsy-based biomarkers for meningioma and brain metastases from NSCLC has been limited in the last five years, as indicated by the PubMed database. Despite this, 2023 saw the publication of a notable study focused on diagnosing NSCLC brain metastases using liquid biopsies. However, this study was unable to develop a model that achieved both high sensitivity and specificity[102]. In the case of meningioma, only one study emerged, yielding promising (84.90% classification accuracy) results examining the methylome of serum/plasma[103].

Our research suggests that the success of our classification is primarily attributable to the use of Raman spectroscopy. This method enabled us to discern potential differences between patient groups across the entire molecular landscape, thereby leading to a robust diagnostic model [43]. For instance, the most pronounced differences for accurate classification between CTRL vs. GBM were observed in bands associated with nucleic acids, as indicated by the vibrations of the phosphate-sugar backbone and phosphate group stretching. Conversely, the differences between CTRL vs. BM were mainly found in the vibrations of lipid acyl chains. Thus, Raman spectroscopy allows for the simultaneous detection of relevant differences across a diverse molecular spectrum.

The diagnostic utility of Raman spectroscopy is further evidenced by other studies. For instance, Shin and colleagues employed plasma EV Raman spectroscopy to detect various types of tumors, achieving a notable 0.956 AUC value[104]. Similarly, Jonak and colleagues effectively differentiated lung cancer from healthy controls with 90% accuracy using linear discriminant analysis[105]. Another study reported over 90% accuracy in detecting breast tumors and predicting surgical outcomes. Li and colleagues also distinguished early-stage pancreatic tumors from controls with a notable 0.950 AUC value[106]. In addition to these examples, several studies have demonstrated the effectiveness of Raman spectroscopic analysis of EVs for the investigation of additional malignancies such as renal[107], prostate[108], ovarian[109] and urogenital cancers[110].

Despite these advances, the application of EVs in diagnosing CNS tumors has been notably absent. To the best of our knowledge, our study represented the first endeavor to classify CNS tumors based on the Raman spectra of sEV-enriched isolates from serum samples. The research to date is proof of concept that for diagnostic purposes, whole-molecule analysis of EVs is appropriate for most tumor types, even CNS tumors, and underlines the importance of considering EVs as a key player in future clinical practice.

5.3. *EVs are carriers of a highly tumor-specific molecular pattern*

The degree of specificity of the molecular content of EVs is elucidated by our Raman spectroscopic analysis conducted on clinical samples, as well as by our meta-analysis examining the proteomes of nine distinct tumor types.

Many studies have shown that analyzing EVs, due to the tumor-related molecular patterns they carry, can be utilized for distinguishing between cancerous and non-cancerous samples or for further categorizing different tumor types based on their characteristics, such as chemosensitivity. For example, Vinik et al. were able to significantly distinguish between control groups and breast cancer patients through the proteins identified in serum-derived EVs[111]. Li et al. explored plasma EVs to classify leukemia patient groups based on varying resistance to imatinib[65]. Choi et al. made distinctions between primary and metastatic colon tumors[112]. Mallawaaratchy et al. focused on identifying aggressive subtypes of glioblastoma[66], and Rontogianni et al. highlighted the capability of proteomic analysis of EVs in differentiating breast cancer subtypes[67]. The diagnostic utility of vesicles in brain tumors has been validated as well. In a study involving mice, Anastasi et al. utilized principal component analysis to distinctly separate the proteomes of control mice from those with glioblastoma multiforme[113].

However, there was limited information on the extent to which tumor types can be distinguished from each other based on the molecular content of EVs. With this in mind, our meta-analysis aimed to distinguish a wide range of tumors with different tissue origin (60 cell lines from 9 tumor types) based on proteomic data.

In a well-written article, which was the source of the NCI-60 proteomic data set, Hurwitz et al. have already demonstrated that some tumor types are distinguishable from the others[68]. However, approaching this valuable dataset with the machine learning based classifier algorithms suggests that the proteomic content carried by cancer EVs is more specific than expected and previously reported.

A certain degree of specificity was already evident when we performed classification based on the 213 proteins shared by the 60 cell lines of all 9 tumor types (core proteome; CA = 49.14%), and primarily associated with protein translation and localization[68].

This classification accuracy further increased when analyzing the entire proteome (CA = 69.10%) of the 60 cell lines and reached its highest value when utilizing a discriminative protein panel (CA = 91.67%).

The high degree of tumor-specific molecular signatures are evidenced not only by *in vitro* results but also by findings from clinical studies. Our model, based on Raman spectroscopy, was successful in distinguishing between tumor types, such as GBM and brain metastases, which often pose challenges in differential diagnosis using conventional methods like CT and MRI[114].

This high level of specificity becomes especially significant when considering the potential development of an EV-based diagnostic platform whose objective is not only to indicate the potential presence of a tumor but also to identify the specific tissue of origin.

5.4. *EVs may represent a promising prognostic value*

Having a deep understanding of tumor invasiveness and proliferation rates is crucial for effectively treating cancer patients, as these factors play a vital role in determining patient survival. Although the molecular information of EVs to estimate these parameters may appear reasonable, considering their role in various cancer-related processes[115–117], the potential of EVs as predictors of tumor cell invasive and proliferative capacity has been unexplored yet.

Several well performed studies have already elucidated that the molecular content of EVs depends on the invasiveness of tumors and may have a prognostic role[118–120], but no attempt has been made to quantify the invasion capacity and proliferation time based on information carried.

As the results of our meta-analysis suggested, a specific subset of the EV proteome provides information about donor cells' proliferation rate and invasion capacity, key factors in tumor progression and metastasis. The predictive invasion and proliferation panel underwent Reactome pathway analysis, revealing biological mechanisms of the predicted effects.

For example, EV proteins in high invasion capacity tumor cell lines may induce HSF1-dependent transactivation, supported by literature data showing HSF1 amplification in aggressive tumors[121,122]. HSF1, a main transactivator of HSPs expression, including HSP60, HSP70, and HSP90, has multiple effects on cancer progression, promoting invasion and metastasis[123]. Conversely, for example, proteins predicting low invasion may downregulate TGF-β signaling, known to function as a tumor promoter by stimulating epithelial-mesenchymal transition (EMT) and promoting metastasis[55]. Or, for another example, Inactivation of TGF-β signaling has been shown to suppress prostate cancer bone metastasis[124].

We assessed the influence of tissue expression of the proliferation and invasion panel on patient survival using the HPA database. Overall, the effects on survival found in the HPA database and the effect of the proteins on invasion and proliferation capacity as determined in our study were consistent in most of the cases.

While alignment with the HPA database may provide an indirect corroboration of our findings, our recent publication in 2023 offers more substantive support to our hypotheses. This clinical study by Dobra *et al.* revealed a significant association between elevated MMP-9 concentration in EVs from GBM patient plasma and decreased survival, reinforcing the prognostic relevance of EVs[125].

In light of recent research, it is evident that the molecular composition of EVs may have a significant impact, extending beyond diagnostic applications to prognostic implications.

5.5. EVs may reflect the characteristic signaling pathways of the tumor

The comprehensive molecular characterization of tumors, including the identification of their disrupted, characteristic pathways, is essential for effective treatment, such as in the identification of appropriate drug targets[126].

This information is usually obtained by extensive examination of tumor tissue and microenvironment; however, tissue biopsy is often limited by invasiveness and risks to patient safety, and in many cases cannot handle intratumoral heterogeneity[126,127]. Consequently, there is a growing necessity to accomplish thorough characterization of particular tumor types through liquid biopsy methods[127].

As recent reviews elucidate, tumor derived-EVs carry specific signaling and metabolic pathway components[128]. This suggests that by examining the molecular content of EVs, we can gain a more comprehensive picture of the tumor's characteristics and its communication with the microenvironment and distant sites.

Our meta-analysis reveals that proteins demonstrating the most significant variance across the nine tumor types could be indicative of pathways specific to each tumor.

For example, matrix-related processes were found to be specifically involved in CNS and kidney tumors in our pathway analysis. In line with our findings, the structure of the collagen matrix has been shown to significantly affect the survival of patients with glioblastoma, and disordered fibers are associated with a worse prognosis[129]. Similar results hold for kidney cancer, where the structure of the collagen matrix predicts tumor grade[130].

NOTCH signaling is distinctive in colon cancers based on the EV proteome, aligning with previous studies showing its importance in colon cancer cell development[131].

We also found a strong association between the leukemia EV proteome and processes related to the RUNX1 transcription factor, known to play a role in hematological malignancies[132].

Additionally, our enrichment study findings are supported by existing literature on melanoma[133], lung[134,135], and ovarian cancer[136,137].

These findings may suggest that EVs could in the future be considered as messengers of specific tumor strategies and could perhaps be used to find drug targets or to help develop personalized medicine, or to understand the underlying biology of different tumor types.

6. CONCLUSION

Numerous studies, as highlighted in our thesis, have shown that EVs are a rich source of tumor information. Our study introduces a Raman spectroscopy approach that could serve as a new, routinely applicable screening tool due to its speed, accuracy, and patient-friendliness. However, establishing a new diagnostic procedure might require the creation of a further standardized, international database. The specificity of the EV proteome and its ability to reveal tumor characteristics underscore the potential role of EVs in previously uncharted clinical applications. Overall, EVs promise to be indispensable players in future clinical practice.

7. NEW FINDINGS

1. Machine learning methods demonstrated that extracellular vesicles carry a highly tumor-specific molecular pattern.
2. The proteomic content of extracellular vesicles may reflect tumor-specific signaling pathways.
3. Protein panels compiled from extracellular vesicles’ proteome can be used to estimate invasiveness and proliferative capacity of donor cells.
4. Machine learning models based on Raman spectroscopy of small extracellular vesicle-enriched isolates from serum are capable of distinguishing brain tumor patients from controls with high efficiency.

8. ACKNOWLEDGEMENT

I would like to express my deepest and most sincere gratitude to my supervisor, **Krisztina Buzás**, for giving me the opportunity to work in her group, for her scientific guidance and her unwavering trust.

I am also grateful to **Péter Horváth**, Director of the Institute of Biochemistry and Head of the Microscopic Image Analysis and Machine Learning Laboratory, for the scientific and infrastructural support of my work. Similarly, I would like to thank **Márta Széll**, Head of the Doctoral School of Interdisciplinary Meidcine for making it possible for me to begin and continue my PhD studies.

My heartfelt thanks go to all the members of our research team - **Gabriella Dobra, Edina Sebestyén, Maria Harmati, Lilla Pintér** and **Tímea Böröczky** - for their countless scientific support and for creating an exceptionally wonderful working environment.

I am thankful to all my co-authors for their support and fruitful collaboration. Special thanks are extended to **Álmos Klekner** for providing of clinical samples and his professional insights. We are grateful to **Zoltán Kónya** for providing the Raman microscope and to **Juan Gomez-Perez** for performing the measurements.

I am grateful for the financial support from the National Research, Development and Innovation Fund, funded by the ÚNKP-20-3-New National Excellence Programme and the EFOP 3.6.3-VEKOP-16-2017-00009 grant.

Lastly, but most importantly, I wish to extend my heartfelt thanks to my **friends**, whose presence made my university years truly joyful and memorable. I am deeply grateful to my **Grandparents**, as well as **Tibor Papp**—my mother's husband—and my siblings, **Katalin** and **Miklós**, along with Miklós's wife, **Eszter**, for their tireless support in numerous aspects of my life.

My heartfelt thanks go to **Tünde** and **Krisztián**, the parents of my beloved **Márk Lehoczki**, for their warm acceptance and support, which have significantly aided me on my path toward achieving my goals. Following this, I express my deepest gratitude to **Márk** himself, for his steadfast trust and love, the cornerstone of my soul throughout this journey.

My utmost and most genuine thanks go to my **Mother**, whose unwavering perseverance and selfless love made it all possible.

9. REFERENCES

1. Yáñez-Mó, M. *et al.* Biological properties of extracellular vesicles and their physiological functions. *J of Extracellular Vesicle* **4**, 27066 (2015).
2. Théry, C. *et al.* Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *J of Extracellular Vesicle* **7**, 1535750 (2018).
3. Buzas, E. I. The roles of extracellular vesicles in the immune system. *Nat Rev Immunol* **23**, 236–250 (2023).
4. Perpiñá-Clérigues, C. *et al.* Lipidomic landscape of circulating extracellular vesicles isolated from adolescents exposed to ethanol intoxication: a sex difference study. *Biol Sex Differ* **14**, 22 (2023).
5. Wen, C. *et al.* Biological roles and potential applications of immune cell-derived extracellular vesicles. *J of Extracellular Vesicle* **6**, 1400370 (2017).
6. Marar, C., Starich, B. & Wirtz, D. Extracellular vesicles in immunomodulation and tumor progression. *Nat Immunol* **22**, 560–570 (2021).
7. Eustes, A. S. & Dayal, S. The Role of Platelet-Derived Extracellular Vesicles in Immune-Mediated Thrombosis. *Int J Mol Sci* **23**, 7837 (2022).
8. Li, Q., Wang, H., Peng, H., Huyan, T. & Cacalano, N. A. Exosomes: Versatile Nano Mediators of Immune Regulation. *Cancers (Basel)* **11**, 1557 (2019).
9. Harmati, M. *et al.* Small extracellular vesicles convey the stress-induced adaptive responses of melanoma cells. *Sci Rep* **9**, 15329 (2019).
10. Raposo, G. & Stahl, P. D. Extracellular vesicles: a new communication paradigm? *Nat Rev Mol Cell Biol* **20**, 509–510 (2019).
11. Nakamura, K. *et al.* Emerging Role of Extracellular Vesicles in Embryo–Maternal Communication throughout Implantation Processes. *IJMS* **21**, 5523 (2020).
12. Ahmad, S., Srivastava, R. K., Singh, P., Naik, U. P. & Srivastava, A. K. Role of Extracellular Vesicles in Glia-Neuron Intercellular Communication. *Front. Mol. Neurosci.* **15**, 844194 (2022).
13. Liu, S.-Y., Liao, Y., Hosseinifard, H., Imani, S. & Wen, Q.-L. Diagnostic Role of Extracellular Vesicles in Cancer: A Comprehensive Systematic Review and Meta-Analysis. *Front Cell Dev Biol* **9**, 705791 (2021).
14. Al-Nedawi, K. *et al.* Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. *Nat Cell Biol* **10**, 619–624 (2008).
15. Skog, J. *et al.* Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nat Cell Biol* **10**, 1470–1476 (2008).

16. Moeng, S. *et al.* Extracellular Vesicles (EVs) and Pancreatic Cancer: From the Role of EVs to the Interference with EV-Mediated Reciprocal Communication. *Biomedicines* **8**, 267 (2020).

17. Hu, D. *et al.* Cancer-associated fibroblasts in breast cancer: Challenges and opportunities. *Cancer Commun (Lond)* **42**, 401–434 (2022).

18. Thuault, S., Ghossoub, R., David, G. & Zimmermann, P. A Journey on Extracellular Vesicles for Matrix Metalloproteinases: A Mechanistic Perspective. *Front Cell Dev Biol* **10**, 886381 (2022).

19. Maacha, S. *et al.* Extracellular vesicles-mediated intercellular communication: roles in the tumor microenvironment and anti-cancer drug resistance. *Mol Cancer* **18**, 55 (2019).

20. Keerthikumar, S. *et al.* Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. *Oncotarget* **6**, 15375–15396 (2015).

21. Ang, H. L. *et al.* Mechanism of epithelial-mesenchymal transition in cancer and its regulation by natural compounds. *Medicinal Research Reviews* **43**, 1141–1200 (2023).

22. Galindo-Hernandez, O., Serna-Marquez, N., Castillo-Sanchez, R. & Salazar, E. P. Extracellular vesicles from MDA-MB-231 breast cancer cells stimulated with linoleic acid promote an EMT-like process in MCF10A cells. *Prostaglandins, Leukotrienes and Essential Fatty Acids* **91**, 299–310 (2014).

23. Pink, R. C., Beaman, E.-M., Samuel, P., Brooks, S. A. & Carter, D. R. F. Utilising extracellular vesicles for early cancer diagnostics: benefits, challenges and recommendations for the future. *Br J Cancer* **126**, 323–330 (2022).

24. Du, S. *et al.* Extracellular vesicles: a rising star for therapeutics and drug delivery. *J Nanobiotechnol* **21**, 231 (2023).

25. Liu, J. *et al.* Extracellular Vesicles in Liquid Biopsies: Potential for Disease Diagnosis. *Biomed Res Int* **2021**, 6611244 (2021).

26. Brocco, D. *et al.* Circulating Cancer Stem Cell-Derived Extracellular Vesicles as a Novel Biomarker for Clinical Outcome Evaluation. *J Oncol* **2019**, 5879616 (2019).

27. Speakman, M. J. & Turnbull, A. R. Passage of a colon 'cast' following resection of an abdominal aortic aneurysm. *Br J Surg* **71**, 935 (1984).

28. Couch, Y. *et al.* A brief history of nearly EV-erything - The rise and rise of extracellular vesicles. *J Extracell Vesicles* **10**, e12144 (2021).

29. Dobra, G. *et al.* Small Extracellular Vesicles Isolated from Serum May Serve as Signal-Enhancers for the Monitoring of CNS Tumors. *IJMS* **21**, 5359 (2020).

30. Brennan, K. *et al.* A comparison of methods for the isolation and separation of extracellular vesicles from protein and lipid particles in human serum. *Sci Rep* **10**, 1039 (2020).

31. Ramos-Zaldívar, H. M. *et al.* Extracellular vesicles through the blood–brain barrier: a review. *Fluids Barriers CNS* **19**, 60 (2022).

32. Theophilou, G. *et al.* Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology. *Expert Review of Molecular Diagnostics* **15**, 693–713 (2015).

33. Menck, K. *et al.* Characterisation of tumour-derived microvesicles in cancer patients' blood and correlation with clinical outcome. *J of Extracellular Vesicle* **6**, 1340745 (2017).

34. Jakobsen, K. R. *et al.* Exosomal proteins as potential diagnostic markers in advanced non-small cell lung carcinoma. *J Extracell Vesicles* **4**, 26659 (2015).

35. Li, Y., Shi, X., Jia, E., Qin, S. & Yu, F. Extracellular vesicle biomarkers for prostate cancer diagnosis: A systematic review and meta-analysis. *Urologic Oncology: Seminars and Original Investigations* **41**, 440–453 (2023).

36. Alegre, E. *et al.* Circulating melanoma exosomes as diagnostic and prognosis biomarkers. *Clinica Chimica Acta* **454**, 28–32 (2016).

37. König, L. *et al.* Elevated levels of extracellular vesicles are associated with therapy failure and disease progression in breast cancer patients undergoing neoadjuvant chemotherapy. *OncoImmunology* **7**, e1376153 (2018).

38. Silva, J. *et al.* Analysis of exosome release and its prognostic value in human colorectal cancer. *Genes Chromosomes & Cancer* **51**, 409–418 (2012).

39. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *IJMS* **20**, 4781 (2019).

40. Wu, W. *et al.* Glioblastoma multiforme (GBM): An overview of current therapies and mechanisms of resistance. *Pharmacol Res* **171**, 105780 (2021).

41. Gollapalli, K. *et al.* Investigation of serum proteome alterations in human glioblastoma multiforme. *Proteomics* **12**, 2378–2390 (2012).

42. Figueroa, J. M. & Carter, B. S. Detection of glioblastoma in biofluids. *Journal of Neurosurgery* **129**, 334–340 (2018).

43. Choi, D. *et al.* The Impact of Oncogenic EGFRvIII on the Proteome of Extracellular Vesicles Released from Glioblastoma Cells. *Molecular & Cellular Proteomics* **17**, 1948–1964 (2018).

44. Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nat Protoc* **11**, 664–687 (2016).

45. Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman spectroscopy. *Chem. Soc. Rev.* **45**, 1958–1979 (2016).

46. Gualerzi, A. *et al.* Raman spectroscopy uncovers biochemical tissue-related features of extracellular vesicles from mesenchymal stromal cells. *Sci Rep* **7**, 9820 (2017).

47. Mehta, K., Atak, A., Sahu, A., Srivastava, S. & C, M. K. An early investigative serum Raman spectroscopy study of meningioma. *Analyst* **143**, 1916–1923 (2018).

48. Pichardo-Molina, J. L. *et al.* Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. *Lasers Med Sci* **22**, 229–236 (2007).

49. Harris, A. T. *et al.* Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. *Head Neck Oncol* **1**, 34 (2009).

50. Park, J. *et al.* Exosome Classification by Pattern Analysis of Surface-Enhanced Raman Spectroscopy Data for Lung Cancer Diagnosis. *Anal. Chem.* **89**, 6695–6701 (2017).

51. Carmicheal, J. *et al.* Label-free characterization of exosome via surface enhanced Raman spectroscopy for the early detection of pancreatic cancer. *Nanomedicine: Nanotechnology, Biology and Medicine* **16**, 88–96 (2019).

52. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* **49**, 107739 (2021).

53. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* **23**, 40–55 (2022).

54. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biol* **20**, 76, s13059-019-1689–0 (2019).

55. Dai, X. & Shen, L. Advances and Trends in Omics Technology Development. *Front Med (Lausanne)* **9**, 911861 (2022).

56. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* **24**, 695–713 (2023).

57. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat Methods* **15**, 399–400 (2018).

58. Lever, J., Krzywinski, M. & Altman, N. Logistic regression. *Nat Methods* **13**, 541–542 (2016).

59. Ley, C. *et al.* Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc* **30**, 753–757 (2022).

60. Mar, J. C. The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond. *Biophys Rev* **11**, 89–94 (2019).

61. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021).

62. Banerjee, J. *et al.* Machine learning in rare disease. *Nat Methods* **20**, 803–814 (2023).

63. Fawagreh, K., Gaber, M. M. & Elyan, E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* **2**, 602–609 (2014).

64. Jiang, P. *et al.* Big data in basic and translational cancer research. *Nat Rev Cancer* **22**, 625–639 (2022).

65. Li, M.-Y. *et al.* Quantitative Proteomic Analysis of Plasma Exosomes to Identify the Candidate Biomarker of Imatinib Resistance in Chronic Myeloid Leukemia Patients. *Front. Oncol.* **11**, 779567 (2021).

66. Mallawaaratchy, D. M. *et al.* Comprehensive proteome profiling of glioblastoma-derived extracellular vesicles identifies markers for more aggressive disease. *J Neurooncol* **131**, 233–244 (2017).

67. Rontogianni, S. *et al.* Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. *Commun Biol* **2**, 325 (2019).

68. Hurwitz, S. N. *et al.* Proteomic profiling of NCI-60 extracellular vesicles uncovers common protein cargo and cancer type-specific biomarkers. *Oncotarget* **7**, 86999–87015 (2016).

69. Weinstein, J. N. Integromic Analysis of the NCI-60 Cancer Cell Lines. *BD* **19**, 11–22 (2004).

70. Weinstein, J. N. Spotlight on molecular profiling: “Integromic” analysis of the NCI-60 cancer cell lines. *Molecular Cancer Therapeutics* **5**, 2601–2605 (2006).

71. Gholami, A. M. *et al.* Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Reports* **4**, 609–620 (2013).

72. Sinha, A., Ignatchenko, V., Ignatchenko, A., Mejia-Guerrero, S. & Kislinger, T. In-depth proteomic analyses of ovarian cancer cell line exosomes reveals differential enrichment of functional categories compared to the NCI 60 proteome. *Biochemical and Biophysical Research Communications* **445**, 694–701 (2014).

73. DeLosh, R. M. & Shoemaker, R. H. Evaluation of Real-Time In Vitro Invasive Phenotypes. in *Metastasis* (ed. Stein, U. S.) vol. 2294 165–180 (Springer US, New York, NY, 2021).

74. Cell Lines in the In Vitro Screen. https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm.

75. The Human Protein Atlas. *The Human Protein Atlas* https://www.proteinatlas.org/.

76. Demšar, J. *et al.* Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* **14**, 2349–2353 (2013).

77. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).

78. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022).

79. MORPHEUS. https://software.broadinstitute.org/morpheus/.

80. Robert, P. & Escoufier, Y. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics* **25**, 257 (1976).

81. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162 (2014).

82. Demšar, J., Leban, G. & Zupan, B. FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data. *Journal of Biomedical Informatics* **40**, 661–671 (2007).

83. Huang, T., Gong, H., Yang, C. & He, Z. ProteinLasso: A Lasso regression approach to protein inference problem in shotgun proteomics. *Computational Biology and Chemistry* **43**, 46–54 (2013).

84. Deutelmoser, H. *et al.* Robust Huber-LASSO for improved prediction of protein, metabolite and gene expression levels relying on individual genotype data. *Briefings in Bioinformatics* **22**, bbaa230 (2021).

85. Fan, J., Shi, S., Qiu, Y., Liu, M. & Shu, Q. Analysis of signature genes and association with immune cells infiltration in pediatric septic shock. *Front. Immunol.* **13**, 1056750 (2022).

86. Chen, D.-L., Cai, J.-H. & Wang, C. C. N. Identification of Key Prognostic Genes of Triple Negative Breast Cancer by LASSO-Based Machine Learning and Bioinformatics Analysis. *Genes* **13**, 902 (2022).

87. Su, Y. *et al.* Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Computers in Biology and Medicine* **145**, 105409 (2022).

88. Jiang, Y. *et al.* Identification of a six-gene prognostic signature for bladder cancer associated macrophage. *Front. Immunol.* **13**, 930352 (2022).

89. Roy, A. & Chakraborty, S. Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety* **233**, 109126 (2023).

90. Zhang, L. *et al.* Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **264**, 120300 (2022).

91. Huang, W. *et al.* Raman spectroscopy and machine learning for the classification of esophageal squamous carcinoma. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **281**, 121654 (2022).

92. Conti, F. *et al.* Raman spectroscopy and topological machine learning for cancer grading. *Sci Rep* **13**, 7282 (2023).

93. Shankar, G. M., Balaj, L., Stott, S. L., Nahed, B. & Carter, B. S. Liquid biopsy for brain tumors. *Expert Review of Molecular Diagnostics* **17**, 943–947 (2017).

94. Taverna, S. *et al.* Exosomes isolation and characterization in serum is feasible in non-small cell lung cancer patients: critical analysis of evidence and potential role in clinical practice. *Oncotarget* **7**, 28748–28760 (2016).

95. Zhi, F. *et al.* A serum 6-miRNA panel as a novel non-invasive biomarker for meningioma. *Sci Rep* **6**, 32067 (2016).

96. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* **4**, 627–635 (2013).

97. Ronvaux, L. *et al.* Liquid Biopsy in Glioblastoma. *Cancers* **14**, 3394 (2022).

98. Khristov, V. *et al.* Tumor-Derived Biomarkers in Liquid Biopsy of Glioblastoma. *World Neurosurgery* **170**, 182–194 (2023).

99. Gatto, L. *et al.* Liquid Biopsy in Glioblastoma Management: From Current Research to Future Perspectives. *The Oncologist* **26**, 865–878 (2021).

100. Saenz-Antoñanzas *et al.* Liquid Biopsy in Glioblastoma: Opportunities, Applications and Challenges. *Cancers* **11**, 950 (2019).

101. Eibl, R. H. & Schneemann, M. Liquid biopsy and glioblastoma. *Exploration of Targeted Anti-tumor Therapy* **4**, 28–41 (2023).

102. Antunes-Ferreira, M. *et al.* Tumor-educated platelet blood tests for Non-Small Cell Lung Cancer detection and management. *Sci Rep* **13**, 9359 (2023).

103. Herrgott, G. A. *et al.* Detection of diagnostic and prognostic methylation-based signatures in liquid biopsy specimens from patients with meningiomas. *Nat Commun* **14**, 5669 (2023).

104. Shin, H. *et al.* Single test-based diagnosis of multiple cancer types using Exosome-SERS-AI for early stage cancers. *Nat Commun* **14**, 1644 (2023).

105. Jonak, S. T. *et al.* Analyzing bronchoalveolar fluid derived small extracellular vesicles using single-vesicle SERS for non-small cell lung cancer detection. *Sens. Diagn.* **2**, 90–99 (2023).

106. Li, J. *et al.* Highly Sensitive Exosome Detection for Early Diagnosis of Pancreatic Cancer Using Immunoassay Based on Hierarchical Surface-Enhanced Raman Scattering Substrate. *Small Methods* **6**, 2200154 (2022).

107. Kamińska, A. *et al.* Raman spectroscopy of urinary extracellular vesicles to stratify patients with chronic kidney disease in type 2 diabetes. *Nanomedicine: Nanotechnology, Biology and Medicine* **39**, 102468 (2022).

108. Osei, E. B. *et al.* Surface-Enhanced Raman Spectroscopy to Characterize Different Fractions of Extracellular Vesicles from Control and Prostate Cancer Patients. *Biomedicines* **9**, 580 (2021).

109. Ćulum, N. M. *et al.* Characterization of ovarian cancer-derived extracellular vesicles by surface-enhanced Raman spectroscopy. *Analyst* **146**, 7194–7206 (2021).

110. Qian, H. *et al.* Diagnosis of urogenital cancer combining deep learning algorithms and surface-enhanced Raman spectroscopy based on small extracellular vesicles. *Spectrochim Acta A Mol Biomol Spectrosc* **281**, 121603 (2022).

111. Vinik, Y. *et al.* Proteomic analysis of circulating extracellular vesicles identifies potential markers of breast cancer progression, recurrence, and response. *Sci. Adv.* **6**, eaba5714 (2020).

112. Choi, D. *et al.* Quantitative proteomics of extracellular vesicles derived from human primary and metastatic colorectal cancer cells. *J of Extracellular Vesicle* **1**, 18704 (2012).

113. Anastasi, F. *et al.* Proteomics analysis of serum small extracellular vesicles for the longitudinal study of a glioblastoma multiforme mouse model. *Sci Rep* **10**, 20498 (2020).

114. Neska-Matuszewska, M., Bladowska, J., Sąsiadek, M. & Zimny, A. Differentiation of glioblastoma multiforme, metastases and primary central nervous system lymphomas using multiparametric perfusion and diffusion MR imaging of a tumor core and a peritumoral zone—Searching for a practical approach. *PLoS ONE* **13**, e0191341 (2018).

115. Liu, Y. *et al.* Tumor Size Still Impacts Prognosis in Breast Cancer With Extensive Nodal Involvement. *Front. Oncol.* **11**, 585613 (2021).

116. Wang, J. *et al.* Prognostic impact of tumor size on patients with neuroblastoma in a SEER -based study. *Cancer Medicine* **11**, 2779–2789 (2022).

117. Narod, S. A. Tumour Size Predicts Long-Term Survival among Women with Lymph Node-Positive Breast Cancer. *Current Oncology* **19**, 249–253 (2012).

118. Shimada, Y. *et al.* Extracellular vesicle-associated microRNA signatures related to lymphovascular invasion in early-stage lung adenocarcinoma. *Sci Rep* **13**, 4823 (2023).

119. Goberdhan, D. C. I. Large tumour-derived extracellular vesicles as prognostic indicators of metastatic cancer patient survival. *Br J Cancer* **128**, 471–473 (2023).

120. Geng, T., Zheng, M., Wang, Y., Reseland, J. E. & Samara, A. An artificial intelligence prediction model based on extracellular matrix proteins for the prognostic prediction and immunotherapeutic evaluation of ovarian serous adenocarcinoma. *Front. Mol. Biosci.* **10**, 1200354 (2023).

121. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2**, 401–404 (2012).

122. Cyran, A. M. & Zhitkovich, A. Heat Shock Proteins and HSF1 in Cancer. *Front. Oncol.* **12**, 860320 (2022).

123. Lauber, K. *et al.* Targeting the heat shock response in combination with radiotherapy: Sensitizing cancer cells to irradiation-induced cell death and heating up their immunogenicity. *Cancer Letters* **368**, 209–229 (2015).

124. Hao, Y., Baker, D. & Ten Dijke, P. TGF-β-Mediated Epithelial-Mesenchymal Transition and Cancer Metastasis. *IJMS* **20**, 2767 (2019).

125. Dobra, G. *et al.* MMP-9 as Prognostic Marker for Brain Tumours: A Comparative Study on Serum-Derived Small Extracellular Vesicles. *Cancers* **15**, 712 (2023).

126. Paananen, J. & Fortino, V. An omics perspective on drug target discovery platforms. *Briefings in Bioinformatics* **21**, 1937–1953 (2020).

127. Russano, M. *et al.* Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. *J Exp Clin Cancer Res* **39**, 95 (2020).

128. Harmati, M., Bukva, M., Böröczky, T., Buzás, K. & Gyukity-Sebestyén, E. The role of the metabolite cargo of extracellular vesicles in tumor progression. *Cancer Metastasis Rev* **40**, 1203–1221 (2021).

129. Pointer, K. B. *et al.* Association of collagen architecture with glioblastoma patient survival. *JNS* **126**, 1812–1821 (2016).

130. Best, S. L. *et al.* Collagen organization of renal cell carcinoma differs between low and high grade tumors. *BMC Cancer* **19**, 490 (2019).

131. Sikandar, S. S. *et al.* *NOTCH* Signaling Is Required for Formation and Self-Renewal of Tumor-Initiating Cells and for Repression of Secretory Cell Differentiation in Colon Cancer. *Cancer Research* **70**, 1469–1478 (2010).

132. Sood, R., Kamikubo, Y. & Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **129**, 2070–2082 (2017).

133. Lau, C., Killian, K. J., Samuels, Y. & Rudloff, U. ERBB4 Mutation Analysis: Emerging Molecular Target for Melanoma Treatment. in *Molecular Diagnostics for Melanoma* (eds. Thurin, M. & Marincola, F. M.) vol. 1102 461–480 (Humana Press, Totowa, NJ, 2014).

134. Xu, L. *et al.* Effects of reduced platelet count on the prognosis for patients with non-small cell lung cancer treated with EGFR-TKI: a retrospective study. *BMC Cancer* **20**, 1152 (2020).

135. Królczyk, G. *et al.* Altered fibrin clot properties in advanced lung cancer: impact of chemotherapy. *J. Thorac. Dis* **10**, 6863–6872 (2018).

136. Nurgalieva, A. K. *et al.* Sodium-dependent phosphate transporter NaPi2b as a potential predictive marker for targeted therapy of ovarian cancer. *Biochemistry and Biophysics Reports* **28**, 101104 (2021).

137. Klemba, A. *et al.* Hypoxia-Mediated Decrease of Ovarian Cancer Cells Reaction to Treatment: Significance for Chemo- and Immunotherapies. *IJMS* **21**, 9492 (2020).

10. SUPPLEMENTARY MATERIALS

Supplementary table 1. ***Gene Ontology Enrichment for the entire proteome***. Table shows the results of the enrichment analysis of a total of 5,908 proteins identified in 60 EV samples. For each Gene Ontology term category, the top 20 results are listed.

Gene Ontology term category	Pathway name	Enrichment FDR[†]	Number of overlapping genes	Genes in the pathway	Fold Enrichment[‡]
Biological processes	Neutrophil degranulation	9.76E-80	357	576	2.520
	Neutrophil mediated immunity	2.73E-80	365	593	2.502
	Neutrophil activation involved in immune response	1.44E-77	357	584	2.485
	Neutrophil activation	1.36E-78	364	597	2.479
	Granulocyte activation	4.03E-78	367	606	2.462
	Leukocyte degranulation	8.37E-78	380	639	2.418
	Myeloid leukocyte mediated immunity	2.36E-79	391	660	2.409
	Myeloid cell activation involved in immune response	4.62E-75	381	652	2.376
	Exocytosis	3.38E-95	560	1027	2.217
	Regulated exocytosis	8.28E-83	492	903	2.215
	Viral process	8.79E-97	612	1158	2.149
	Intracellular protein transport	3.34E-80	548	1066	2.090
	Intracellular transport	9.58E-126	918	1857	2.010
	Protein transport	3.92E-106	882	1867	1.921
	Establishment of protein localization	4.61E-110	932	1989	1.905
	Secretion by cell	4.64E-77	690	1489	1.884
	Export from cell	1.03E-78	713	1545	1.876
	Cellular protein localization	2.14E-100	901	1960	1.869

Gene Ontology term category	Pathway name	Enrichment FDR[†]	Number of overlapping genes	Genes in the pathway	Fold Enrichment[‡]
	Cellular macromolecule localization	6.66E-101	907	1974	1.868
	Secretion	1.28E-77	741	1636	1.841
Molecular functions	Cadherin binding	1.76E-83	262	354	3.009
	Cell adhesion molecule binding	1.87E-102	387	577	2.727
	Structural constituent of ribosome	3.53E-31	130	201	2.630
	Structural molecule activity	4.08E-68	421	784	2.183
	GTPase activity	2.44E-26	170	319	2.167
	GTP binding	5.49E-27	204	407	2.038
	Actin binding	9.55E-31	237	477	2.020
	RNA binding	3.96E-127	925	1873	2.008
	Nucleoside-triphosphatase activity	5.57E-46	361	735	1.997
	Guanyl nucleotide binding	3.98E-26	210	429	1.990
	Guanyl ribonucleotide binding	3.98E-26	210	429	1.990
	Hydrolase activity, acting on acid anhydrides	6.59E-48	389	802	1.972
	Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	6.59E-48	389	802	1.972
	Pyrophosphatase activity	3.39E-47	386	798	1.967
	Protein-containing complex binding	2.77E-78	678	1446	1.906
	Cytoskeletal protein binding	1.27E-49	479	1050	1.855
	Oxidoreductase activity	2.44E-26	348	835	1.694
	Adenyl nucleotide binding	2.46E-40	676	1743	1.577
	ATP binding	3.20E-37	641	1662	1.568

Gene Ontology term category	Pathway name	Enrichment FDR†	Number of overlapping genes	Genes in the pathway	Fold Enrichment‡
	Adenyl ribonucleotide binding	8.55E-39	667	1730	1.568
Cellular components	Primary lysosome	2.55E-40	126	170	3.013
	Focal adhesion	4.23E-89	323	471	2.788
	Cell-substrate junction	4.23E-89	326	478	2.773
	Secretory granule lumen	4.82E-48	224	368	2.475
	Vesicle lumen	2.32E-48	227	374	2.468
	Cytoplasmic vesicle lumen	1.28E-47	225	372	2.459
	Anchoring junction	1.74E-93	511	907	2.291
	Collagen-containing extracellular matrix	8.41E-43	247	446	2.252
	Ribonucleoprotein complex	5.13E-73	432	792	2.218
	Secretory granule	1.49E-76	507	975	2.114
	Secretory vesicle	5.93E-73	563	1150	1.990
	Lytic vacuole	5.22E-51	417	865	1.960
	Lysosome	5.22E-51	417	865	1.960
	Perinuclear region of cytoplasm	5.10E-44	373	783	1.937
	Vacuole	2.72E-50	453	973	1.893
	Vesicle membrane	8.14E-59	619	1405	1.791
	Cytoplasmic vesicle membrane	1.17E-56	607	1385	1.782
	Endosome	5.93E-41	495	1168	1.723
	Organelle envelope	2.74E-42	558	1352	1.678
	Envelope	**2.74E-42**	558	1352	1.678

†FDR (False Discovery Rate) is calculated based on nominal P-value from the hypergeometric test. FDR tells us how likely the enrichment is by chance.
‡Fold Enrichment is defined as the percentage of genes in our list belonging to a pathway, divided by the corresponding percentage in the human genome.

Supplementary table 2. ***Gene Ontology Enrichment for the core proteome.*** The table shows the enrichment analysis of the 213 proteins shared by all EV samples. For each Gene Ontology term category, the top 20 results are listed.

Gene Ontology term category	Pathway name	Enrichment FDR[†]	Number of overlapping genes	Genes in the pathway	Fold Enrichment[‡]
Biological processes	SRP-dependent cotranslational protein targeting to membrane	9.54E-29	27	112	26.807
	Cotranslational protein targeting to membrane	3.50E-28	27	118	25.444
	Establishment of protein localization to endoplasmic reticulum	1.18E-29	29	132	24.430
	Protein targeting to ER	1.31E-28	28	128	24.325
	Protein localization to endoplasmic reticulum	1.79E-28	30	162	20.593
	Cytoplasmic translation	5.78E-28	30	169	19.740
	MRNA catabolic process	8.07E-30	44	458	10.683
	Establishment of protein localization to organelle	2.95E-32	50	566	9.823
	RNA catabolic process	2.19E-28	44	501	9.766
	Regulated exocytosis	8.18E-30	57	903	7.019
	Viral process	1.62E-36	71	1158	6.818
	Protein localization to organelle	3.90E-31	62	1040	6.629
	Exocytosis	1.38E-30	61	1027	6.605
	Intracellular protein transport	1.37E-30	62	1066	6.468
	Cellular protein localization	8.07E-36	87	1960	4.936
	Cellular macromolecule localization	9.37E-36	87	1974	4.901
	Establishment of protein localization	7.88E-34	85	1989	4.752
	Cell activation	5.89E-27	70	1658	4.695

Gene Ontology term category	Pathway name	Enrichment FDR[†]	Number of overlapping genes	Genes in the pathway	Fold Enrichment[‡]
	Intracellular transport	3.25E-30	78	1857	4.671
	Protein transport	2.31E-29	77	1867	4.586
Molecular functions	RNA binding	8.93E-40	89	1873	5.284
	Hydrolase activity, acting on acid anhydrides	4.84E-33	57	802	7.903
	Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	4.84E-33	57	802	7.903
	Pyrophosphatase activity	4.84E-33	57	798	7.943
	Structural molecule activity	1.38E-26	50	784	7.092
	Nucleoside-triphosphatase activity	1.08E-34	57	735	8.624
	Cell adhesion molecule binding	2.62E-27	45	577	8.672
	Guanyl nucleotide binding	1.33E-19	33	429	8.554
	Guanyl ribonucleotide binding	1.33E-19	33	429	8.554
	ATP hydrolysis activity	3.42E-14	27	415	7.235
	GTP binding	2.69E-19	32	407	8.743
	Ubiquitin-like protein ligase binding	7.02E-13	24	359	7.434
	Cadherin binding	5.29E-29	39	354	12.251
	Ubiquitin protein ligase binding	2.37E-13	24	341	7.826
	GTPase activity	3.82E-20	30	319	10.458
	Structural constituent of ribosome	3.98E-22	27	201	14.937
	Unfolded protein binding	5.33E-17	20	139	16.000
	Structural constituent of cytoskeleton	1.79E-13	16	113	15.745
	GDP binding	5.40E-18	17	74	25.546

Gene Ontology term category	Pathway name	Enrichment FDR[†]	Number of overlapping genes	Genes in the pathway	Fold Enrichment[‡]
	Protein folding chaperone	1.46E-16	14	47	33.123
Cellular components	Cytosolic small ribosomal subunit	6.46E-18	16	62	28.697
	Cytosolic ribosome	4.70E-28	27	123	24.410
	Melanosome	4.70E-28	27	123	24.410
	Pigment granule	4.70E-28	27	123	24.410
	Small ribosomal subunit	8.68E-16	17	100	18.904
	Ficolin-1-rich granule lumen	8.94E-21	23	143	17.885
	Focal adhesion	1.22E-50	61	471	14.402
	Ribosomal subunit	1.72E-22	28	218	14.283
	Cell-substrate junction	1.53E-50	61	478	14.191
	Ficolin-1-rich granule	1.52E-17	24	224	11.914
	Ribosome	2.37E-20	29	291	11.082
	Secretory granule lumen	1.12E-21	33	368	9.972
	Cytoplasmic vesicle lumen	1.44E-21	33	372	9.865
	Vesicle lumen	1.57E-21	33	374	9.812
	Anchoring junction	1.23E-41	68	907	8.337
	Secretory granule	5.70E-26	54	975	6.159
	Ribonucleoprotein complex	2.95E-19	42	792	5.897
	Secretory vesicle	7.58E-27	59	1150	5.705
	Supramolecular complex	2.61E-15	50	1455	3.821
	Synapse	8.70E-15	49	1443	3.776

[†] FDR (False Discovery Rate) is calculated based on nominal P-value from the hypergeometric test. FDR tells us how likely the enrichment is by chance.

[‡] Fold Enrichment is defined as the percentage of genes in our list belonging to a pathway, divided by the corresponding percentage in the human genome.

Supplementary table 3. ***The selected 172 proteins.*** The table contains the 172 proteins selected from the entire proteome with LASSO algorithm.

Tumor type	Gene symbols
Breast	KIAA1324, RAB27B, ENPP1, EMD, PVRL2, PHLDA1, KIF11, TBCC, MUC1, UBQLN2, AFF4, UBTD2, MGP, NCAM2, TRAPPC2L, STX6, SLK, JAK1, ARMC6, VANGL1, LTBP1, EPHB4, IFIT1
CNS	IGFBP5, DKK3, CA9, LOXL1, RCOR3, COL1A2, RCN2, THY1, EIF2B3, STEAP3, CSRP2, GPM6A, KCTD12, SPARC, GPX2, ATP6V1H, MYL2, SERPINB1, PLEKHB2, CPOX, FAT1, LOXL4, SUMO1, ALDH2, CLTB, HIST1H2BK, DPP4, VSNL1, CLDN12, SPAST, ABCB1, TOM1L1
Colon	DPP7, MUC13, AGO2, CGN, EREG, PALD1, NAA25, CSTF3, CLDN4, SLC1A1, VIL1, COPS5, ST14, MYCBP, CMBL, LTBP3, TRAP1, EFEMP2, PDLIM1, CDCP1, TPP1, SMAD5
Kidney	GAL3ST1, CCBE1, CYB5A, OCIAD2, SQRDL, IL4I1, GLS, MGLL, TSKU, GLB1, PAPLN, CD70, TRAPPC2L, NPC1
Leukemia	MOB3A, CORO7, PPP6R1, SMARCD2, PTPRC, MYO1G, VRK1, LCP1, KDM1A, CD53, CORO1A, ICAM3, ARID1A, SMARCA4, METTL1, CLEC11A, DSC1, C3, ABCB1
Lung	FGA, FGG, DMBT1, HIST1H2AH, MUC5B, MUC5AC, ARG1, LTA4H, PTGR1, GPD2, AKR1B10, TMED7, AKR1D1, VWF
Melanoma	PMEL, ITIH5, HSPA2, CTHRC1, CSPG4, NDUFA8, GPNMB, LOXL3, MRPL41, VGF, LAMC2, FBLN2, LSM1, PCNP, MAN1A2, SPON1
Ovary	MFAP5, GDA, SLC20A2, CA2, RAB11FIP1, SLC34A2, TAGLN, APH1A, HK2, TSPAN11
Prostate	NEFM, TNXB, PPP1R14B, TNFRSF10A, LRRC8A, LARP4B, TMF1, ASPH, MST1R, FGFBP1, UBQLN1, PRKCI, ECI1, FAM84B, LAMA3

Supplementary table 4. ***Comparison of the invasion panel with the Human Protein Atlas database*** (number of patient for breast = 1075, CNS = 153, colon = 597, kidney = 877, lung = 994, melanoma = 102, ovary = 373, prostate = 494).

		Data obtained from Human Proteome Atlas						
Predictive state in our study (+: positively correlates with invasion, -: negatively correlates with invasion)	**Gene symbol**	**Prognostic in**	**Difference in 5-year survival rate for high expression compared to low expression**	**5-year survival for low expression**	**5-year survival for high expression**	**FPKM cut off**[†]	***p***[‡]	**Consistency with our study**
+	CAV2	CNS	-1.00%	10%	9%	2.75	4.20E-03	Consistent
+		Colon	-10.00%	66%	56%	2.1	4.60E-02	Consistent
+		Kidney	-11.00%	75%	64%	18.26	3.40E-03	Consistent
+		Lung	-5.00%	48%	43%	8.96	2.30E-02	Consistent
+	COL11A1	Breast	-10.00%	87%	77%	10.95	2.60E-03	Consistent
+		Kidney	-22.00%	80%	58%	0.07	3.00E-12	Consistent
+		Lung	-9.00%	52%	43%	0.89	2.20E-03	Consistent
+		Prostate	-3.00%	100%	97%	0.03	1.90E-02	Consistent
+	CRYAB	CNS	10.00%	7%	17%	541.4	2.70E-02	Not consistent
+		Colon	-26.00%	69%	43%	3.59	4.90E-05	Consistent
+	DNAJB4	Kidney	-14.00%	79%	65%	4.93	2.30E-04	Consistent
+	EDIL3	Breast	-9.00%	86%	77%	4.89	9.90E-03	Consistent
+		Colon	18.00%	48%	66%	1.58	2.70E-02	Not consistent
+		Kidney	-13.00%	79%	66%	2.86	1.00E-03	Consistent
+		Melanoma	-61.00%	61%	0%	1.22	6.60E-03	Consistent

		Data obtained from Human Proteome Atlas						
Predictive state in our study (+: positively correlates with invasion, -: negatively correlates with invasion)	**Gene symbol**	**Prognostic in**	**Difference in 5-year survival rate for high expression compared to low expression**	**5-year survival for low expression**	**5-year survival for high expression**	**FPKM cut off[†]**	***p*[‡]**	**Consistency with our study**
+	THY1	CNS	-13.00%	13%	0%	49.86	9.60E-04	Consistent
+		Kidney	-15.00%	78%	63%	13.19	2.40E-06	Consistent
+		Lung	-5.00%	47%	42%	13.11	2.40E-02	Consistent
+		Ovary	-15.00%	36%	21%	44.44	2.90E-03	Consistent
+	VCAN	Kidney	-12.00%	77%	65%	7.6	5.20E-04	Consistent
+		Lung	-7.00%	49%	42%	8.08	1.20E-02	Consistent
+		Ovary	-17.00%	38%	21%	7.02	4.40E-03	Consistent
-	CAPN7	Breast	-10.00%	85%	75%	9.09	2.70E-02	Not consistent
-		CNS	8.00%	4%	12%	5.39	2.10E-02	Consistent
-		Colon	28.00%	56%	84%	6.79	4.80E-05	Consistent
-		Kidney	12.00%	61%	73%	3.93	1.10E-03	Consistent
-	FIS1	CNS	-6.00%	13%	7%	74.5	1.30E-02	Not consistent
-		Lung	6.00%	41%	47%	34.2	2.00E-05	Consistent
-		Kidney	17.00%	62%	79%	50.36	1.10E-07	Consistent
-		Prostate	3.00%	97%	100%	66.16	3.50E-02	Consistent
-		Breast	13.00%	73%	86%	32.43	1.50E-02	Consistent
-	HIST1H3A	CNS	-15.00%	20.00%	5%	0.11	3.90E-02	Not consistent
-		Colon	13.00%	58.00%	71%	0.22	4.60E-02	Consistent

Predictive state in our study (+: positively correlates with invasion, -: negatively correlates with invasion)	Gene symbol	Data obtained from Human Proteome Atlas						Consistency with our study
		Prognostic in	Difference in 5-year survival rate for high expression compared to low expression	5-year survival for low expression	5-year survival for high expression	FPKM cut off†	$p^‡$	
-		Kidney	7.00%	67.00%	74%	0.18	8.80E-03	Consistent
-		Lung	8.00%	42.00%	50%	0.15	4.00E-02	Consistent
-		Ovary	4.00%	30.00%	34%	0.55	4.40E-02	Consistent
-	MYO18A	Colon	28.00%	57.00%	85.00%	10.33	4.00E-02	Consistent
-		Kidney	11.00%	65.00%	76%	6.48	3.20E-04	Consistent
-		Lung	-14.00%	55.00%	41.00%	3.77	3.30E-02	Not consistent
-	PARP1	Colon	27.00%	39%	66%	17.55	7.40E-03	Consistent
-		Kidney	-19.00%	74%	55%	15.99	1.30E-05	Not consistent
-	PPP1CC	CNS	4.00%	7%	11%	23.67	1.70E-02	Consistent
-		Colon	5.00%	58%	63%	24.22	4.00E-02	Consistent
-		Kidney	-11.00%	72%	61%	21.1	2.90E-03	Not consistent
-	PSME3	Lung	-9.00%	51.00%	42%	21.47	6.00E-03	Not consistent
-		Colon	17.00%	49.00%	66%	25.16	1.60E-03	Consistent
-		Kidney	10.00%	61.00%	71%	12.52	5.00E-02	Consistent
-		Ovary	-7.00%	34.00%	27%	25.25	1.30E-02	Not consistent
-		Lung	-13.00%	52.00%	39.00%	9.25	1.30E-02	Not consistent
-	RGS19	Colon	-24.00%	67.00%	43.00%	6.53	4.30E-03	Not consistent
-		Kidney	-19.00%	76.00%	57.00%	7.81	2.10E-08	Not consistent

		Data obtained from Human Proteome Atlas						
Predictive state in our study (+: positively correlates with invasion, -: negatively correlates with invasion)	**Gene symbol**	**Prognostic in**	**Difference in 5-year survival rate for high expression compared to low expression**	**5-year survival for low expression**	**5-year survival for high expression**	**FPKM cut off[†]**	***p*[‡]**	**Consistency with our study**
-		Breast	5.00%	78.00%	83.00%	6.5	4.00E-02	Consistent
-		Melanoma	25.00%	28.00%	53.00%	8.49	2.40E-02	Consistent
-	SLC43A3	CNS	-18.00%	21.00%	3.00%	6.76	1.00E-02	Not consistent
-		Kidney	-25.00%	79.00%	54.00%	5.94	4.00E-13	Not consistent
-		Prostate	3.00%	96.00%	99.00%	1.12	4.00E-02	Consistent
-	SYNGR2	Breast	11.00%	77%	88%	55.18	2.20E-02	Consistent
-		Kidney	-13.00%	74.00%	61%	44.98	5.40E-04	Not consistent
-		Melanoma	62.00%	0%	62%	20.18	1.30E-02	Consistent
-		Ovary	4.00%	30%	34%	75.12	1.20E-02	Consistent
-	TGFBRAP1	CNS	7.00%	7%	14%	7.1	2.60E-02	Consistent
-	VPS11	CNS	8.00%	6%	14%	12.28	2.00E-02	Consistent
-		Kidney	12.00%	62%	74%	13.01	1.40E-04	Consistent

[†] The FPKM (fragments per kilobase per million) cut off is a default value defined in the Human Protem Atlas.

[‡] The p value is the result of a log-rank test, which can be found in the Human Proteome Atlas database.

Supplementary 5. ***Comparison of the proliferative capacity panel with the Human Protein Atlas database*** (number of patient for breast = 1075, CNS = 153, colon = 597, kidney = 877, lung = 994, melanoma = 102, ovary = 373, prostate = 494).

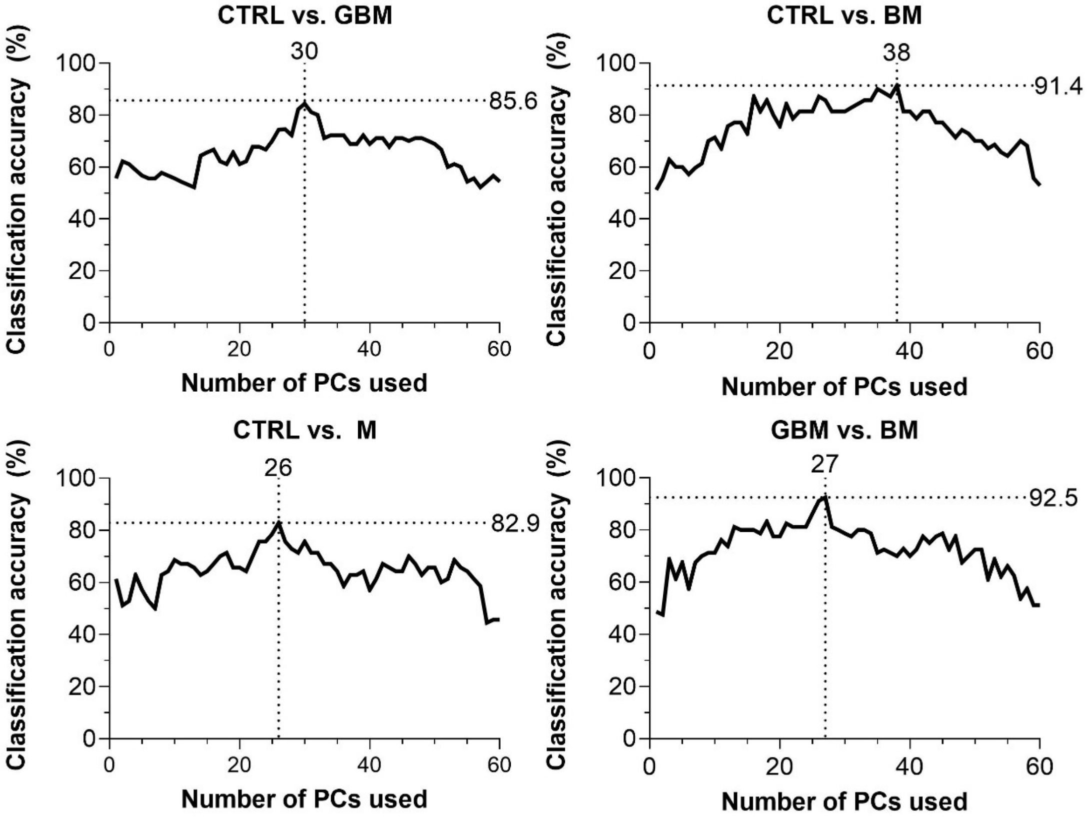
Predictive state for doubling time (hrs) in our study (+: positively correlates with dobling time, -: negatively correlates with doubling time)	**Gene symbol**	**Data obtained from Human Proteome Atlas**						**Consistency with our study**
		Prognostic in	**Difference in 5-year survival rate for high expression compared to low expression**	**5-year survival for low expression**	**5-year survival for high expression**	**FPKM cut off**[†]	***p***[‡]	
+	ACOT13	Ovary	0.26	0.12	0.38	3.78	2.50E-06	Consistent
+	ECH1	CNS	0.14	0.04	0.18	38.61	8.00E-03	Consistent
+	ERGIC1	Colon	0.18	0.5	0.68	14.71	3.40E-02	Consistent
+		Lung	-0.19	0.56	0.37	10.82	7.00E-03	Not consistent
+	FKBP2	Breast	0.03	0.8	0.83	25.8	4.90E-03	Consistent
+		Kidney	0.05	0.68	0.73	68.82	1.70E-02	Consistent
+		Ovary	0.11	0.27	0.38	54.71	5.90E-03	Consistent
+	GLUD1	CNS	0.12	0	0.12	52.26	5.10E-04	Consistent
+		Kidney	0.18	0.58	0.76	61.83	1.40E-06	Consistent
+		Ovary	-0.18	0.45	0.27	26.53	4.60E-03	Not consistent
+	PDZK1IP1	CNS	-0.15	0.15	0	1.08	9.50E-04	Not consistent
+		Kidney	0.11	0.66	0.77	485.31	7.90E-03	Consistent
+		Prostate	0.02	0.97	0.99	14.02	2.60E-02	Consistent
+	ZBC1D2	Breast	-0.05	0.83	0.78	8.78	2.30E-02	Not consistent
+		CNS	-0.12	0.12	0	3.08	2.40E-02	Not consistent
+		Colon	0.14	0.52	0.66	6.75	1.10E-02	Consistent

Predictive state for doubling time (hrs) in our study (+: positively correlates with dobling time, -: negatively correlates with doubling time)	Gene symbol	Data obtained from Human Proteome Atlas						Consistency with our study
		Prognostic in	Difference in 5-year survival rate for high expression compared to low expression	5-year survival for low expression	5-year survival for high expression	FPKM cut off†	$p^{\ddagger}$	
+		Ovary	-0.14	0.39	0.25	5.34	9.20E-03	Not consistent
-	ANAPC1	Kidney	-0.1	0.76	0.66	0.81	2.10E-02	Consistent
-		Melanoma	-0.42	0.58	0.16	1.52	3.90E-02	Consistent
-	DSG2	CNS	-0.22	0.25	0.03	0.07	3.00E-04	Consistent
-		Colon	0.26	0.48	0.74	58.05	7.80E-04	Consistent
-		Lung	-0.09	0.5	0.41	35.92	2.40E-04	Consistent
-		Melanoma	-0.44	0.73	0.29	0.24	5.00E-02	Consistent
-	FRYL	CNS	0.09	0.07	0.16	3.92	4.90E-02	Not consistent
-		Colon	0.08	0.57	0.65	4.58	6.70E-03	Not consistent
-		Lung	-0.04	0.46	0.42	3.23	4.80E-02	Consistent
-		Melanoma	-0.13	0.47	0.34	1.96	4.70E-02	Consistent
-	HDAC1	Breast	0.07	0.77	0.84	26.22	2.50E-02	Not consistent
-		Kidney	-0.17	0.78	0.61	20.22	1.00E-05	Consistent
-		Melanoma	-0.21	0.55	0.34	21.89	5.00E-02	Consistent
-		Ovary	-0.09	0.34	0.25	38	3.50E-02	Consistent
-		Prostate	-0.04	1	0.96	41.2	4.40E-02	Consistent
-	SLC38A1	Breast	-0.04	0.83	0.79	58.87	8.90E-03	Consistent
-		Kidney	0.14	0.58	0.72	18.99	1.20E-02	Not consistent

Predictive state for doubling time (hrs) in our study (+: positively correlates with dobling time, -: negatively correlates with doubling time)	**Gene symbol**	**Data obtained from Human Proteome Atlas**						**Consistency with our study**
		Prognostic in	**Difference in 5-year survival rate for high expression compared to low expression**	**5-year survival for low expression**	**5-year survival for high expression**	**FPKM cut off[†]**	***p*[‡]**	
-		Lung	-0.05	0.48	0.43	10.94	2.00E-03	Consistent
-		Ovary	-0.12	0.37	0.25	19.84	4.20E-03	Consistent
-	SLC38A5	Kidney	-0.3	0.76	0.46	1.35	2.20E-16	Consistent
-		Ovary	-0.12	0.37	0.25	1.61	3.80E-03	Consistent
-	THOC2	Breast	-0.07	0.84	0.77	9.56	2.10E-02	Consistent
-		CNS	0.08	0.06	0.14	4.81	1.80E-02	Not consistent
-		Kidney	-0.11	0.77	0.66	3.82	4.30E-03	Consistent
-		Melanoma	-0.44	0.44	0	7.42	1.00E-02	Consistent
-		Ovary	0.15	0.25	0.4	4.89	2.20E-02	Not consistent
-	WDR12	Breast	-0.05	0.83	0.78	4.09	1.10E-02	Consistent
-		Melanoma	-0.62	0.62	0	4.44	4.90E-02	Consistent
-		Ovary	0.11	0.25	0.36	3.2	4.80E-02	Not consistent

[†] The FPKM (fragments per kilobase per million) cut off is a default value defined in the Human Protem Atlas.

[‡] The *p* value is the result of a log-rank test, which can be found in the Human Proteome Atlas databa



Supplementary figure 1. ***Receiver Operating Characteristic (ROC) curves for the classification models.*** Intersecting black dotted lines show sensitivity, specificity and corresponding threshold values of the group-membership score, with black filled circles at their intersections. (BM: brain metastasis of non-small cell lung cancer, CTRL: control, lumbar disc hernia, GBM: glioblastoma multiforme, M: meningioma, PC: principal component)

cancers

MDPI

Article

Raman Spectral Signatures of Serum-Derived Extracellular Vesicle-Enriched Isolates May Support the Diagnosis of CNS Tumors

Matyas Bukva [1,2], Gabriella Dobra [1,2], Juan Gomez-Perez [3], Krisztian Koos [1], Maria Harmati [1], Edina Gyukity-Sebestyen [1], Tamas Biro [4,5], Adrienn Jenei [6], Sandor Kormondi [7], Peter Horvath [1], Zoltan Konya [3], Almos Klekner [6] and Krisztina Buzas [1,8,*]

[1] Laboratory of Microscopic Image Analysis and Machine Learning, Biological Research Centre, Institute of Biochemistry, Eötvös Loránd Research Network (ELKH), H-6726 Szeged, Hungary; bukva.matyas@brc.hu (M.B.); dobra.gabriella@brc.hu (G.D.); koos.krisztian@brc.hu (K.K.); harmati.maria@brc.hu (M.H.); sebestyen.edina@brc.hu (E.G.-S.); horvath.peter@brc.hu (P.H.)

[2] Department of Medical Genetics, Doctoral School of Interdisciplinary Medicine, University of Szeged, H-6720 Szeged, Hungary

[3] Department of Applied and Environmental Chemistry, University of Szeged, H-6720 Szeged, Hungary; juan.gomez@chem.u-szeged.hu (J.G.-P.); konya@chem.u-szeged.hu (Z.K.)

[4] Department of Immunology, Faculty of Medicine, University of Debrecen, H-4032 Debrecen, Hungary; biro.tamas@med.unideb.hu

[5] Monasterium Laboratory, D-48149 Münster, Germany

[6] Clinical Centre, Department of Neurosurgery, University of Debrecen, H-4032 Debrecen, Hungary; jenei.adrienn@med.unideb.hu (A.J.); klekner.almos@med.unideb.hu (A.K.)

[7] Department of Traumatology, University of Szeged, H-6720 Szeged, Hungary; kormondi.sandor.pal@med.u-szeged.hu

[8] Department of Immunology, University of Szeged, H-6720 Szeged, Hungary

* Correspondence: buzas.krisztina@brc.hu; Tel.: +36-62-432-340

check for updates

Citation: Bukva, M.; Dobra, G.; Gomez-Perez, J.; Koos, K.; Harmati, M.; Gyukity-Sebestyen, E.; Biro, T.; Jenei, A.; Kormondi, S.; Horvath, P.; et al. Raman Spectral Signatures of Serum-Derived Extracellular Vesicle-Enriched Isolates May Support the Diagnosis of CNS Tumors. *Cancers* **2021**, *13*, 1407. https://doi.org/10.3390/cancers13061407

Academic Editors: Brigitta G. Baumert and Janusz Rak

Received: 22 January 2021
Accepted: 16 March 2021
Published: 19 March 2021

Simple Summary: The conventional central nervous system (CNS) tumor diagnostic methods, especially the invasive intracranial surgical tissue sample collecting, imposes a heavy burden on both patients and healthcare providers. We aimed to explore the potential role of serum-derived small extracellular vesicles (sEVs) in diagnosing CNS tumors through Raman spectroscopic analyses. A relevant number of clinical samples (138) were obtained from four patient groups, namely glioblastoma multiforme, brain metastasis of non-small-cell lung cancer, meningioma, and lumbar disc herniation as controls. After the isolation and Raman measurements of sEV-sized particles, the Principal Component Analysis–Support Vector Machine algorithm was performed on the Raman spectra for pairwise classifications. The groups compared were distinguishable with 80–95% sensitivity and 80–90% specificity. Our results support that Raman spectroscopic analysis of sEV-sized particles is a promising liquid-biopsy-based method that could be further developed in order to be applicable in the diagnosis of CNS tumors.

Abstract: Investigating the molecular composition of small extracellular vesicles (sEVs) for tumor diagnostic purposes is becoming increasingly popular, especially for diseases for which diagnosis is challenging, such as central nervous system (CNS) malignancies. Thorough examination of the molecular content of sEVs by Raman spectroscopy is a promising but hitherto barely explored approach for these tumor types. We attempt to reveal the potential role of serum-derived sEVs in diagnosing CNS tumors through Raman spectroscopic analyses using a relevant number of clinical samples. A total of 138 serum samples were obtained from four patient groups (glioblastoma multiforme, non-small-cell lung cancer brain metastasis, meningioma and lumbar disc herniation as control). After isolation, characterization and Raman spectroscopic assessment of sEVs, the Principal Component Analysis–Support Vector Machine (PCA–SVM) algorithm was performed on the Raman spectra for pairwise classifications. Classification accuracy (CA), sensitivity, specificity and the Area Under the Curve (AUC) value derived from Receiver Operating Characteristic (ROC) analyses were used to evaluate the performance of classification. The groups compared were distinguishable with

82.9–92.5% CA, 80–95% sensitivity and 80–90% specificity. AUC scores in the range of 0.82–0.9 suggest excellent and outstanding classification performance. Our results support that Raman spectroscopic analysis of sEV-enriched isolates from serum is a promising method that could be further developed in order to be applicable in the diagnosis of CNS tumors.

Keywords: small extracellular vesicles; Raman spectroscopy; glioblastoma multiforme; brain metastasis; meningioma; CNS tumors; liquid biopsy

1. Introduction

In recent decades, secreted extracellular vesicles (EVs) have been recognized as a pathway for intercellular communication in both eukaryotes and prokaryotes [1]. Furthermore, several studies demonstrate the role of EVs in maintaining cellular homeostasis and integrity by compensating for the stress condition [2–4]. Their involvement in different pathophysiological processes has already been highlighted, especially in malignant diseases [5,6]. EVs released by tumor cells are involved in both stromal and distant cell communication, metastatic niche formation, and immune cell suppression [7–13].

Recent clinical research has highlighted that EVs could serve as novel tools for various therapeutic approaches, including oncotherapy, vaccination, immune-modulatory or regenerative therapies, and drug delivery [14]. Furthermore, EVs are gaining increasing popularity in biomarker research as their potential in liquid biopsy has been recognized [15].

Secreted EVs are stably present in body fluids, and represent a concentrated sample of the cytosolic milieu (proteins, nucleic acids, and lipids) of the donor cells [16–18]. It has been shown that EVs isolated from the serum and plasma offer a useful tool to improve the signal-to-noise ratio in analytics by assuring to abundant protein depletion (such as albumin and lipoproteins) and enriching the tumor-specific molecular composition [19,20]. Moreover, EVs can cross various biological barriers, such as the blood–brain barrier (BBB), and easily enter the peripheral blood [21,22].

Examining the protein, nucleic acid, or lipid contents of EVs has revealed several molecules as promising diagnostic markers for different tumor types. For example, glypican-1 glycoprotein enriched in circulating EVs has been shown to be suitable for distinguishing malignant pancreatic cancer from benign malformations with 100% classification accuracy [23,24].

Given their favorable biological properties, serum-derived EVs are being evaluated in the diagnosis and monitoring of central nervous system (CNS) tumors which represent a major challenge in oncology [25].

Today, the diagnosis of CNS tumors mainly relies on neuroimaging techniques (e.g., magnetic resonance imaging (MRI) or computer tomography (CT)) and tissue biopsy. However, all of these methods have numerous limitations [26]. Among others, MRI can only detect tumor masses of sufficient magnitude, and has little prognostic value in terms of long-term recurrence [27]. Distinguishing between different CNS malignancies, such as glioblastoma multiforme and brain metastases, is also challenging using neuroimaging techniques [28]. In addition, treatment-related changes can overlap with residual or recurrent tumors, making tumor monitoring highly challenging [29].

Many brain tumors are particularly difficult to be sampled or are inaccessible for tissue biopsy. Even in cases of biopsy, the procedure harbors significant risks for the patient (e.g., hemorrhage, or impairment of neurological functions). These risks and difficulties hamper not only the diagnosis, but also the monitoring of treatment response or distinguishing tumor recurrence from pseudoprogression [30]. In addition, in some cases, such as in glioblastoma multiforme, the focal sampling of small and localized tumor tissues may not fully capture intratumoral heterogeneity [31,32].

Liquid biopsy has remarkable advantages over conventional methods, offering a minimally invasive, safer, faster, and cheaper way to diagnose and monitor malignant diseases. Tumor tissues release various types of biomarkers, such as proteins, nucleic acids or lipids, and EVs that accumulate in body fluids (including the blood, urine, cerebrospinal fluid and saliva) are accessible for sampling [33–35].

Tumor markers determined from blood samples, such as the prostate-specific antigen (PSA), alpha-fetoprotein and cancer antigen 125 (CA-125), have already been introduced into clinical practice to support the diagnosis and/or monitoring of prostate, liver and ovarian cancers. Research is underway to identify other non-invasive biomarkers for the monitoring of a broader range of malignant diseases [36].

However, identifying blood-based CNS tumor markers is more challenging, presumably explained by several reasons. BBB can prevent tumor-derived molecules (tumor "information") from entering the peripheral blood, therefore molecules released by other tissues/cells at high concentrations can impede the detection of potential tumor biomarkers present in lower concentrations. Abundant serum proteins (such as albumin or lipoproteins) also appear as a significant analytical noise [19,37]. Nevertheless, investigations for blood-based CNS tumor markers are in the spotlight of neuro-oncological research, as they would have outstanding advantages in patient care [38].

Due to their beneficial properties detailed above, EVs are promising tools in the research for CNS tumor biomarkers. Several published studies aimed to examine the nucleic acid and protein contents of blood samples or EVs derived from CNS tumor patients (specifically, from patients with gliomas). However, these studies attempted to identify one or two biomarkers of nucleic acid or protein types, and these molecules did not prove to be sufficiently specific or sensitive to serve as diagnostic markers, and thus they were not validated with blinded samples [39–41].

Analyzing the whole molecular composition of tumor-related EVs isolated from blood samples may provide a solution to overcome the difficulties encountered in CNS tumor biomarker research. Raman spectroscopy is a suitable approach for this purpose, since it provides information on the entire molecular content of a sample. Raman spectroscopy is a non-destructive, label-free vibrational technique that measures the non-elastic scattering effect induced by a radiating laser. The energy of this inelastically scattered light is reduced by the vibrational energy of the chemical bonds present in the molecules within a sample [42]. The difference is proportional to, and thus specifically refers to, the chemical composition of the sample. Therefore Raman spectroscopy can reveal a specific spectral signature that describes the whole chemical composition, and thus avoids the need for identifying any specific protein, nucleic acid, or lipid biomarkers [43]. In addition, Raman spectroscopy may be suitable for the characterization of EVs by identifying different subtypes by origin and function, which is an important and long-standing challenge for EV-based biomarker research [44].

Recent studies suggest that Raman spectroscopic analysis of the whole molecular composition of various sample types may be suitable to develop promising diagnostic methods for clinical practice [45–47]. Some of these in vitro studies focused on EVs and demonstrated their outstanding diagnostic efficiency. For example, using this technique, Parks and colleagues distinguished EVs released by lung cancer cells from those secreted by normal cells with 95.3% sensitivity and 97.3% specificity. Meanwhile, discriminatory spectral differences were also identified using principal component analysis (PCA) [48]. Charmichael and colleagues revealed that EVs originating from pancreatic cancer cells were distinguishable from those released by normal pancreatic epithelial cells with 90% accuracy [49].

However, no studies to date have investigated the diagnostic efficiency of Raman spectroscopic analysis of serum-derived EVs with regard to CNS tumors. Thus, we aimed to explore the potential role of serum-derived EVs in diagnosing CNS tumors through Raman spectroscopic analyses on a relevant number of clinical samples.

For this purpose, 138 serum samples were obtained from four patient groups. The serum samples were collected from patients diagnosed with the two most common types of brain tumors, namely malignant glioblastoma multiforme (GBM) and the typically benign meningioma (M), as well as from patients with a prevalent brain metastasis originating from non-small-cell lung cancer (BM). Patients with lumbar disc herniation without evidence of neurological cancer served as controls (CTRL) [20,50,51]. Particles within the size range of small EVs (sEVs) were isolated from serum samples via differential centrifugation, and were assessed by a Raman microscope. Multivariate analyses, including Principal Component Analysis–Support Vector Machine (PCA–SVM) and FreeViz, as well as conventional statistical methods, such as Receiver Operating Characteristic (ROC) analysis, were carried out on spectroscopic data to develop and evaluate a classification model.

Our results support that analyzing the serum-derived sEV-enriched isolates by Raman spectroscopy, which captures the whole molecular composition, may be suitable to develop a method with a possible diagnostic value for CNS tumors, and thus it may have the potential to be introduced into clinical practice in the future.

2. Results

2.1. Particles Isolated from Serum Show sEV Properties

Particles were isolated by differential centrifugation from 138 serum samples of patients with GBM, BM), M and CTRL. Isolated sEVs were characterized by transmission electron microscopy (TEM) and nanoparticle tracking analysis (NTA), as well as by examining characteristic sEV markers (Alix, CD81 and calnexin) by Western blotting (WB) (Figure 1). Average concentration, mean and mode diameter of the particles were measured as 7.41 × 10^{10} particles/mL, 111.20 nm and 83.32 nm, respectively. Alix, CD81 positivity and calnexin negativity was determined (see Figure S1 for the original WB images).

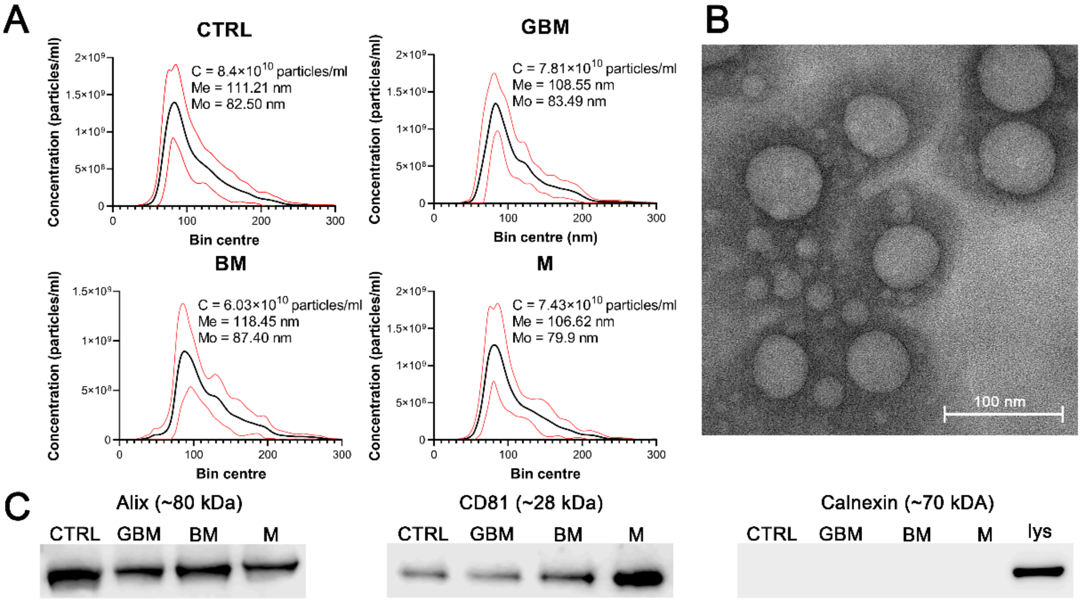


Figure 1. Characterization of the particles. The figure represents the results of the particle characterization: size distribution of the sEV samples isolated from the four patient groups (black and red lines represent the mean and the standard deviation of the concentration, respectively) (**A**), a representative TEM image of the sEVs (**B**), and the Western blot analysis of the sEV markers (**C**). (Abbreviations: CTRL, control; GBM, glioblastoma multiforme; BM, brain metastasis; M, meningioma; C, particle concentration; lys, cell lysate; Me, mean diameter size; Mo, mode diameter size.)

No statistically significant differences were identified among the patient groups in any of the parameters of the isolated particles.

2.2. Patient Groups Can Be Distinguished Using the PCA–SVM Algorithm with High Classification Efficiency

Raman spectroscopic analyses of the isolated 138 samples yielded 5 spectra per sample. The spectral range between 801 cm^{-1} and 3100.5 cm^{-1} was investigated. After standard normal variate (SNV) normalization and PCA transformation, the classification of samples was performed using the SVM algorithm. Classification efficiency was evaluated by classification accuracy (CA), sensitivity, specificity and the area under the curve (AUC) value derived from the ROC analysis. Relevant spectral differences were revealed by PCA. Figure 2 shows the flowchart of Raman spectroscopy data processing.

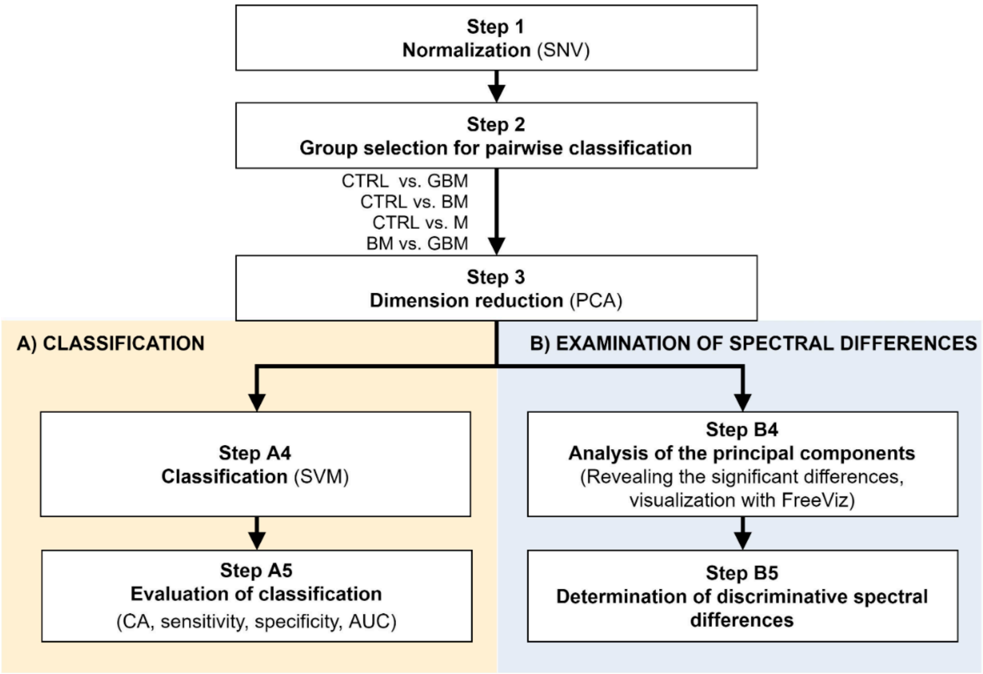

Step 1
Normalization (SNV)
Step 2
Group selection for pairwise classification
CTRL vs. GBM
CTRL vs. BM
CTRL vs. M
BM vs. GBM
Step 3
Dimension reduction (PCA)
A) CLASSIFICATION
B) EXAMINATION OF SPECTRAL DIFFERENCES
Step A4
Classification (SVM)
Step A5
Evaluation of classification
(CA, sensitivity, specificity, AUC)
Step B4
Analysis of the principal components
(Revealing the significant differences, visualization with FreeViz)
Step B5
Determination of discriminative spectral differences

Figure 2. Workflow of Raman spectroscopic data processing. The figure shows the analysis step by step. After Step 3, the workflow separates (parts **A** and **B**) according to the purpose of the analysis. (Abbreviations: AUC, area under the curve; CA, classification accuracy; SNV, standard normal variate; SVM, support-vector machine; PCA, principal component analysis).

After averaging the spectra, row normalization was performed using the SNV method (Step 1) (Figure 3).

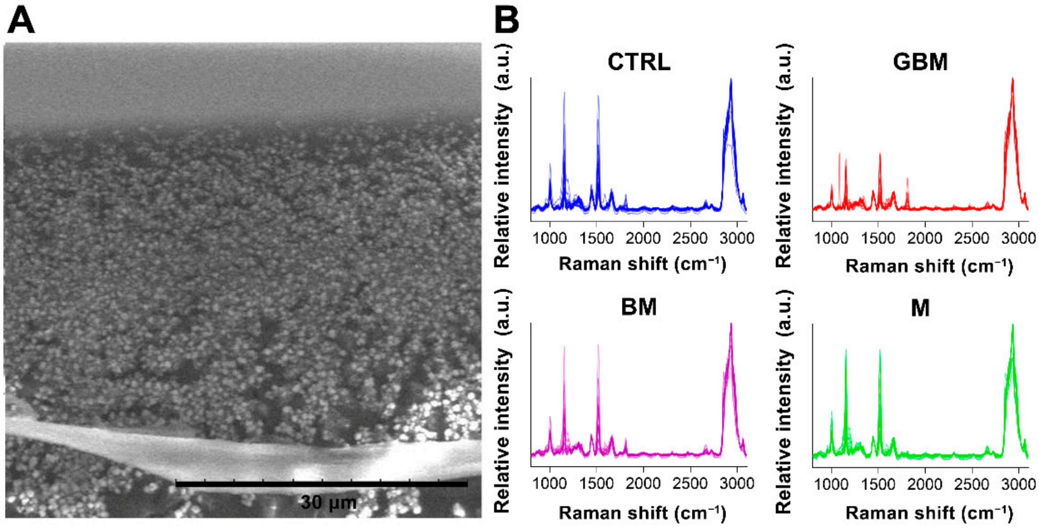


Figure 3. Particles on a calcium fluoride substrate (scanning electron microscope image) (**A**). Averaged and SNV-normalized spectra of the four patient groups (**B**). (Abbreviations: a.u., arbitrary unit).

Following SNV-normalization, the spectra for the samples of the four patient groups were compared pairwise (each patient group was compared to the control, and BM vs. GBM was compared) for two purposes: first, to develop and test a classification algorithm, and second, to identify relevant spectral differences. PCA applied on the pairwise comparisons reduced multivariate data dimensions by transforming the original variables (wavenumbers) into a smaller number of new variables, i.e., principal components (PCs) (Step 3).

Pairwise comparisons were conducted using the linear SVM (Step A4) algorithm, yielding classification models for each paired group. To make predictions for the test samples, a minimum threshold for the group-membership score was determined. Test samples with scores above this threshold were classified into the target group of interest. The optimal score thresholds were automatically set to correspond to the highest classification accuracy (CA, the ratio of correctly classified samples per all samples).

CA was 85.6% for CTRL vs. GBM, 91.4% for CTRL vs. BM, 82.9% for CTRL vs. M and 92.5% for BM vs. GBM. The best classification performance was achieved when a certain number of PCs were included in the models: 30 PCs for CTRL vs. GBM, 38 PCs for CTRL vs. BM, 27 PCs for CTRL vs. M, and 26 PCs for BM vs. GBM (Figure 4).

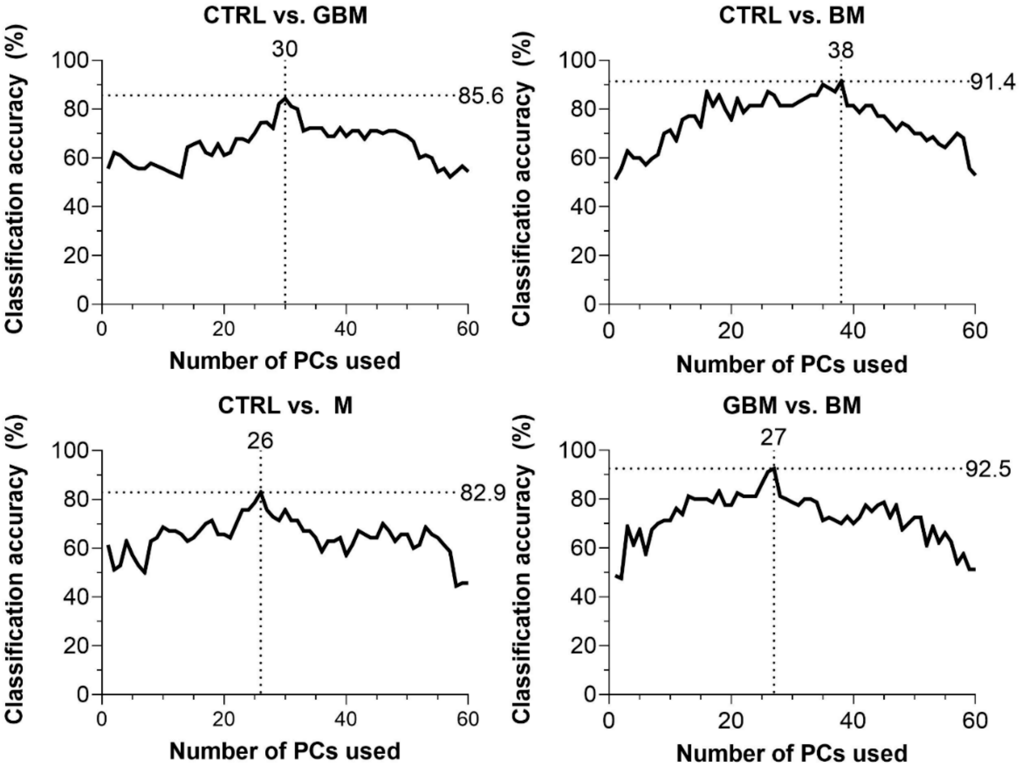


Figure 4. Classification accuracy (CA) scores with respect to the number of PCs included in the model (60 PCs at a maximum). Black dotted lines show the highest CA peaks with the corresponding number of PCs. (Abbreviations: PC, principal component).

Sensitivity and specificity were evaluated as further metrics of classification performance. ROC analyses of the pairwise classification models yielded four graphs showing the automatically set optimal thresholds (having the highest CA value), with related sensitivity, specificity and AUC values, as well as *p*-value (Figure 5).

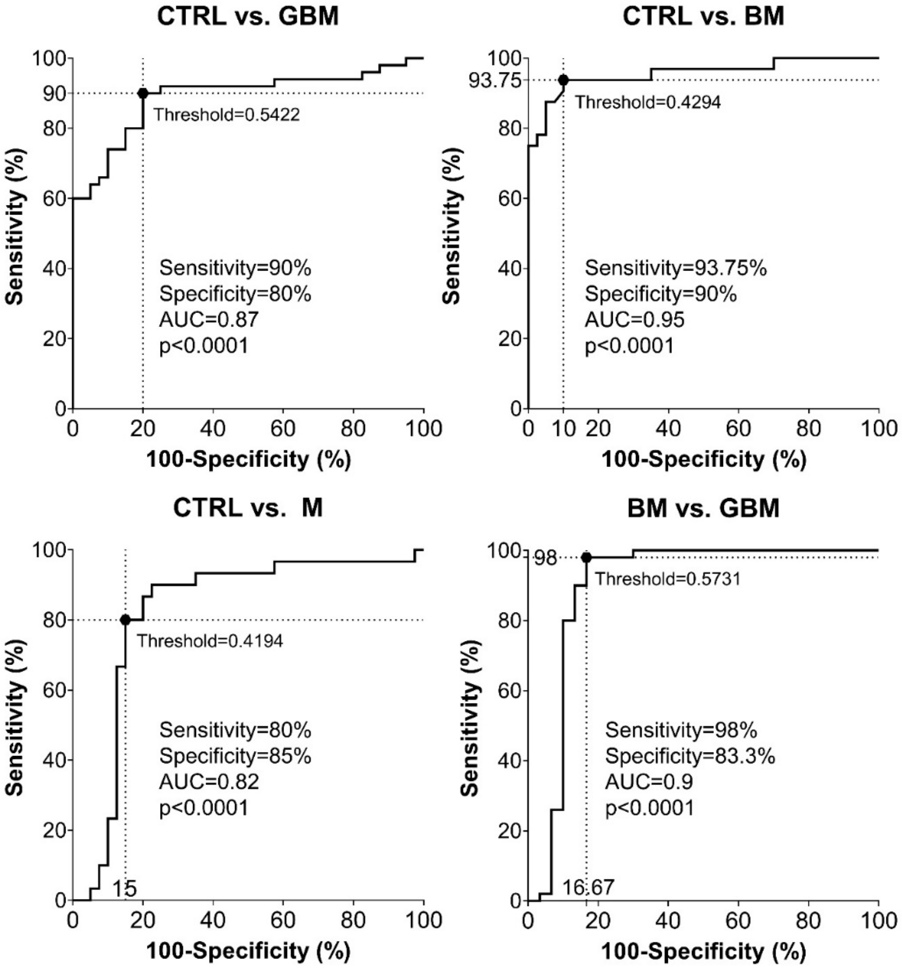


Figure 5. Receiver Operating Characteristic (ROC) curves for the classification models. Intersecting black dotted lines show sensitivity, specificity and corresponding threshold values of the group-membership score, with black filled circles at their intersections.

As shown in the graphs in Figure 5, using the optimal thresholds, the classification models were able to distinguish GBM, BM and M patients from CTRL patients with a sensitivity and specificity of 90% and 80%, 93.75% and 90%, 80% and 85%, respectively (Step A5). Using the classification model, the two malignancies, BM and GBM, could be distinguished from each other with a sensitivity of 98% and a specificity of 83.3%. In the same order of pairwise comparisons (GBM, BM and M patients vs. CTRL, and BM vs. GBM), the AUC values were 0.87, 0.95, 0.82 and 0.9, respectively ($p < 0.0001$ in all cases).

2.3. *Analysis of the PCs Revealed Discriminative Spectral Differences*

Next, differences in the molecular content of serum-derived sEV-enriched isolates from each group were investigated to reveal the spectral differences relevant with regard to the classification. SNV-normalized spectra and the PCs obtained from PCA were analyzed using the FreeViz method, in order to reveal and visualize relevant spectral differences.

The FreeViz method (Step B4) displayed the optimized projections of the multivariate data sets in a 2-dimensional scatterplot (Figure 6). Based on the length and direction of PC vectors, two PCs that were revealed to play the most important role in distinguishing each paired group (marked with a yellow background in Figure 6) were further assessed to determine discriminative spectral signatures.

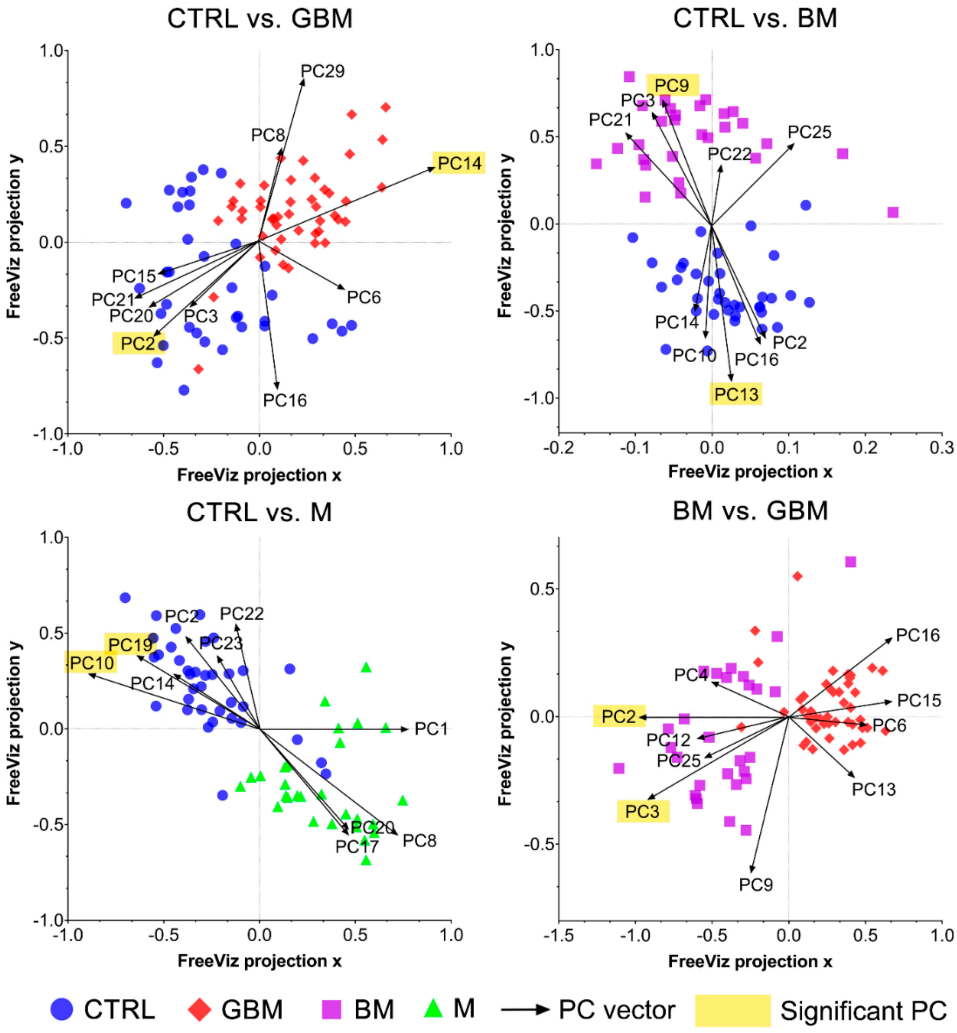


Figure 6. FreeViz projections of pairwise comparisons. Analysis of the PCA-transformed data using the FreeViz method yielded four graphs. Different dots and colors represent the patient groups and healthy controls. Black vectors represent the PCs. In each graph, only the 10 most relevant PC vectors were plotted. For each comparison, PCs marked with a yellow background indicate the 2 most significant PCs.

Based on the results of the FreeViz method and p-values from Welch's *t*-test, PC14 and PC2, PC9 and PC13, P10 and PC19, and PC2 and PC3 explained most of the discriminative differences in the CTRL vs. GBM, CTRL vs. BM, CTRL vs. M and BM vs. GBM comparisons, respectively ($p < 0.05$ in all cases) (see Figure S2 for the score plots of the selected PCs).

Evaluating the selected PCs, we attempted to find the chemical bonds and functional groups corresponding to the spectral differences found to have an important role in distinguishing the compared groups (Step B5).

Regarding the CTRL vs. GBM comparison, most of the discriminative spectral differences were characteristic for carbohydrates, such as bands associated with a pyranose ring (800–975 cm^{-1}), O-H deformation vibrations (1030–1080 cm^{-1}) and C-O stretching vibrations (1030–1290 cm^{-1}). These bands largely overlap with the region's characteristic for nucleic acids, including the bands associated with the vibrations of the phosphate-sugar backbone (800–1000 cm^{-1}), symmetric and asymmetric phosphate group stretching vibrations (1000–1250 cm^{-1}), glycosidic bond vibrations (1250–1550 cm^{-1}), and in-plane double bond vibrations of bases (1530–1780 cm^{-1}) (Figure 7).

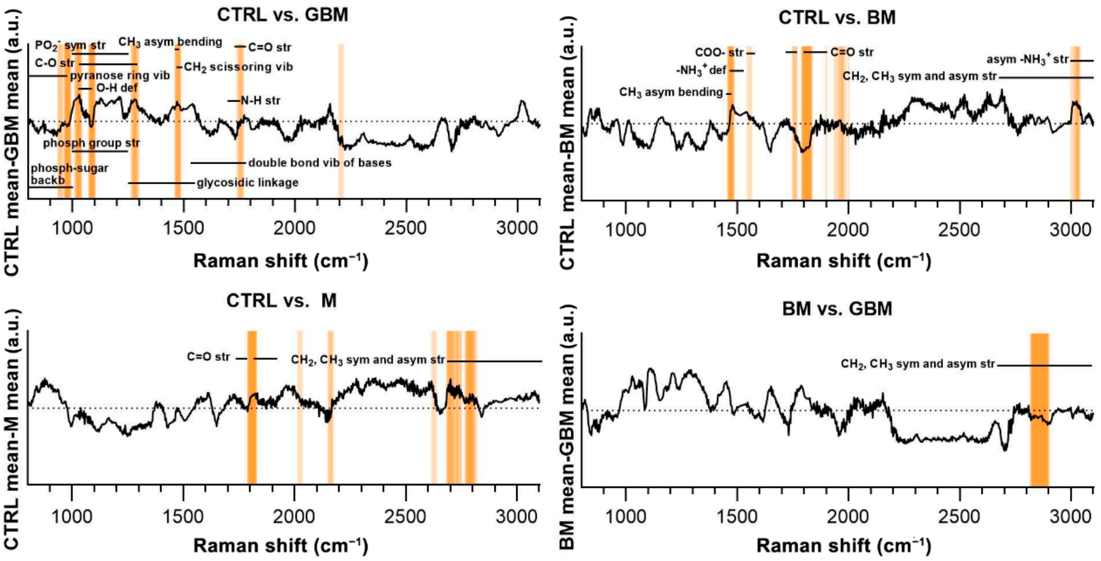


Figure 7. Subtraction spectra for the pairwise comparisons. Subtraction spectra were produced by subtracting the mean signal intensities for the groups compared. Spectral regions having a higher-than-average contribution to significant PCs were marked with orange bars. The more saturated a bar is, the more that region is represented on the selected PCs. The dotted horizontal line represents zero difference at y = 0. (Abbreviations: asym, asymmetric; backb, backbone; def, deformation; phosph, phosphate; str, stretching; sym, symmetric; vib, vibration).

Regarding the CTRL vs. BM comparison, the wavenumbers found to have an important role in distinguishing the BM group from the control mainly correlated with lipids (CH_3 asymmetrical bending (1470–1490 cm^{-1}), CH_2 and CH_3 symmetrical and asymmetrical stretching vibrations (2700–3100 cm^{-1})) and amino acids (–NH_3^+ deformation band (1485–1150 cm^{-1}), –NH_3^+ asymmetrical stretching (3000–3100 cm^{-1}), carboxylate ion stretching (1560–1600 cm^{-1}) and C=O stretching vibrations of the carboxyl group (1700–1755 cm^{-1})). Regarding the CTRL vs. M and BM vs. GBM comparisons, the wavenumbers highly correlated with vibrations originating from acyl chains of lipids, such as CH_3 and CH_2 symmetric and asymmetric stretching vibrations (2700–3100 cm^{-1}) (see Table S1 for the tabular form of the discriminative spectral differences).

3. Discussion

Circulating sEVs are considered as promising sources of CNS tumor markers. Several studies have investigated the nucleic acid and protein contents of blood samples or EVs from CNS tumor patients. These studies have generally attempted to identify one or two biomarkers targeting the proteome, genome or lipidome. However, these molecules alone do not have sufficient diagnostic or prognostic value, thus they cannot be used as single biomarkers, and none of these have been validated on blinded clinical samples [26,39–41,52,53].

Analyzing the entire molecular composition of tumor-related EVs could provide a solution to overcome the difficulties encountered in CNS tumor biomarker research. Raman spectroscopy is a suitable approach for this purpose, as it provides information on the total molecular content, yielding a specific spectral signature that describes the chemical composition of a sample. Thus, it has the potential to avoid the need for identifying any specific protein, nucleic acid or lipid biomarkers [43].

Based on these considerations, we have attempted to explore the potential role of serum-derived sEVs in the diagnosis of CNS tumors through Raman spectroscopic analyses on a clinically relevant cohort. According to our knowledge, this is the first study that aims to classify CNS tumors based on the Raman spectra of sEV-enriched isolates from serum samples.

For this purpose, 138 serum samples obtained from four patient groups were analyzed. Serum samples were collected from three brain tumor groups considered as the most common malignant, benign and metastatic brain tumors (GBM, BM, M) and from a control group (CTRL) [20,50,51]. sEV-sized particles from serum samples were isolated by differential centrifugation.

The particles found in the isolates show sEV properties (CD81, Alix positivity and calnexin negativity). However, since abundant serum protein aggregates and lipoproteins (LPs) are able to mimic sEVs in terms of size (mean and mode diameter of 111.20 nm and 83.32 nm), we cannot state that only sEVs are present in the isolates. In our previously published proteomic-based study on the same patient groups, we have shown that, although contaminants are still present in the isolates, differential centrifugation significantly enriched the sEV-specific markers and reduced the level of LPs. Since LPs and abundant serum protein aggregates are certainly present in addition to sEVs, the isolates should be considered only as sEV-enriched rather than purified sEVs. In light of these, Raman spectra may characterize a circulating particle profile, part of which is sEVs assumed as biomarkers.

No significant differences were found between the four patient groups in the concentration, mean and mode diameter of sEV-sized particles. Osti et al. observed higher EV concentration in the plasma samples of GBM patients, brain metastases and extra-axial brain tumors compared to healthy controls [54]. Other researchers also showed higher EV concentration in tumor patients when unfractionated EV isolates or a broader spectrum of EVs were analyzed [55–57]. However, other non-neoplastic diseases of the central nervous system can also increase the number of sEVs, as has been shown in acute ischemic stroke or multiple sclerosis patients [58,59]. These findings suggest that the elevated sEV concentration cannot be clearly attributed to the presence of the tumor as immune responses or other systemic responses also contribute to the circulating EV population. Therefore, the intense inflammation associated with lumbar disc herniation (CTRL) may explain why no statistical difference was identified between tumorous and non-tumorous patient groups [60].

In the light of the above, we hypothesize that the isolates contain not only tumor tissue-derived vesicles but also other circulating vesicles, including vesicles released by red blood cells, platelets and immune cells. Therefore, the differences observed in the Raman spectra of the different patient groups may not only reflect tumor-specific processes but other host responses, i.e., the tumor-associated immune responses or different coagulant phenotypes as well [61–63].

After the Raman spectroscopic measurements, multivariate analyses and conventional statistical methods were applied on the spectroscopic data to develop and evaluate a classification model, and find the characteristic spectral signatures distinguishing between the patient groups and healthy controls, as well as between the glioblastoma multiforme and brain metastasis groups.

PCA was applied to all the SNV-normalized spectra. PCA is a standard way to reduce data dimensionality and obtain characteristic spectral signatures [48,64].

Classification was performed by applying the SVM algorithm on PCA-transformed data. Classification performance was evaluated by CA, sensitivity (rate of true positive samples), specificity (rate of true negative samples), and the AUC value derived from ROC analysis, which are all commonly used and accepted metrics in clinical practice.

The GBM, BM and M groups proved to be distinguishable from CTRL with 85.6%, 91.4%, 82.9% of CA, respectively. Interestingly, maximal classification accuracy depended on the number of PCs used for classification, showing an increasing trend towards a specific number of PCs (Figure 4). The relationship between the number of PCs included and CA achieved is probably explained by the complexity of these biological samples.

In most studies, the first two PCs (PC1 and PC2) were able to describe the complete data set and revealed distinctive patterns [64]. However, as Lyng and colleagues' findings show, the first two components may not sufficiently explain the information included in the complete Raman spectrum for biological samples, due to their complex molecular composition [65]. Using combinations of PCA and various discriminant analyses, Lyng and colleagues found that 20 PCs were required to separate breast tumor tissue samples from healthy controls with 80% CA. Our results also support that including two PCs, only one cannot develop an accurate classification model capable of spectrally discriminating between different ex vivo biological sample groups. However, classification performance can be improved by increasing the number of PCs included in the model, although above a certain number of PCs used, the information they explain may be meaningless or may account for noise, leading to decreased classification accuracy (Figure 4). This suggests that the spectra for biological samples show a high degree of overlap due to their complexity. Hence, accurate classification can be performed only when one correctly uses several dimensions, taking small spectral differences into account.

Although sensitivity and specificity can be calculated by regarding each value of the group-membership scores as a threshold, the ROC curves, including all possible decision thresholds, plus AUC together, offer a more comprehensive assessment [66,67].

According to the ROC analyses, sensitivity and specificity values were as follows: 90% and 80% for CTRL vs. GBM, 93.75 and 90% for CTRL vs. BM, 80% and 85% for CTRL vs. M, and 98% and 83.3% for BM vs. GBM, respectively. In the same order of comparisons, AUC values were 0.87, 0.95, 0.82 and 0.9 (Figure 5).

Based on the literature of ROC analysis, our classification models for CTRL vs. GBM and CTRL vs. M comparisons can be considered as "excellent", and "outstanding" for CTRL vs. BM and BM vs. GBM comparisons [66].

Due to its reliable theoretical basis, SVM has become one of the most widely used classification methods in recent years, especially for complex multivariate data sets obtained from spectroscopic analyses, characterized by high variance and probable outliers [65,68–70]. These properties make the SVM classifier particularly suitable to discriminate between clinical samples based on their Raman spectra, even for diseases known to be highly heterogeneous (such as GBM) [31,32]. Furthermore, Neska-Matuszewska highlighted that various malignancies (e.g., BM and GBM) are challenging to be distinguished using conventional neuroimaging techniques [28]. In light of our findings, Raman spectra-based SVM classification may support a reliable differential diagnosis between primary brain tumors and metastatic brain malignancies. However, it should be noted that the future confirmation of our results via the comparison of other primary and metastatic brain tumor types is clearly required.

Using the FreeViz method, PCs that were particularly important in terms of distinguishing between the compared groups could be identified (Figure 6). By examining the contribution of the wavenumbers to the selected PCs, we attempted to find the chemical bonds and functional groups that correlate with the spectral differences revealed to play an important role in our pairwise classifications.

The molecular correlation of the vibrational bands in the Raman spectra is extremely difficult to interpret. The difficulty arises from the complexity of biological samples in which an abundance of organic molecules coexist and share some of the functional groups responsible for the Raman-spectral features [71]. As a result, the overlap of different vibrational bands hinders the precise identification of any specific molecules based on Raman spectral features (Figure 7). Nevertheless, it was possible to identify discriminative spectral differences in the CTRL vs. GBM comparison, defining bands characteristic for carbohydrates and nucleic acids. These differences may be due to the characteristic metabolism of GBM, as it is associated with a significant increase in glycolysis for energy production and abnormal purine and pyrimidine synthesis [72]. Comparing the spectra of the CTRL and BM groups, significant differences were found in the characteristic bands of lipids and amino acids, which can be partly explained by the fact that an NSCLC appears to be reliant on fatty acid and serine catabolism [73]. Comparing the CTRL and M groups, as well as the two malignant groups BM and GBM, the lipid bands had outstanding importance with regard to discriminatory differences. The prominent importance of lipids in the BM vs. GBM comparison may be explained by the increased lipid catabolism of NSCLC and the elevated level of de novo lipid synthesis in GBM [72,73]. However, more detailed identification based on the difficulties described above is not expedient for complex biological samples.

It should be noted that co-purification of abundant serum proteins and LP particles in EV isolation methods is a common and well-known challenge [74]. Liu and colleagues emphasized that serum is not the perfect choice for representative sampling of circulating EVs, as a high proportion of EVs may be lost during clotting, and blood components enrolled in the coagulation may also (e.g., platelets) release EVs altering the original content of blood samples [58]. Some cancerous diseases, such as GBM, may also have a procoagulant phenotype [75].

Despite these difficulties, we have revealed in a previously published article that EV isolation from the serum samples of the same patient groups significantly improves the signal-to-noise ratio, even in the case of GBM with an elevated procoagulant activity [20]. Although abundant serum proteins and LPs were still present in EV isolates, isolation depleted their concentration and enriched the EV and tumor-specific protein markers. These results are consistent with previous similar researches on serum-derived EVs [19,76]. Nevertheless, examining plasma instead of serum should be considered in further investigations [58,74].

Enciso-Martinez and colleagues have determined Raman spectral signatures which were able to distinguish EVs from LPs and platelets with 95% confidence [77]. These special signature regions were found at 1004 cm^{-1} and between 2811 cm^{-1} and 3023 cm^{-1}. Wavelength 1004 cm^{-1} had a strong peak in EVs but was not present in LPs and platelets. Furthermore, in the spectral range of 2811–3023 cm^{-1}, EVs showed stronger intensity after 2900 cm^{-1} ('protein component of the CH region') compared to the spectra of LPs, where the region before 2900 cm^{-1} ('lipid component of the CH region') proved to be more intense.

The Raman spectra of sEV-enriched isolates in our study show similar properties: a peak with strong intensity is present at 1004 cm^{-1} and the "protein component" of the CH region was found to be much more prominent than the "lipid component".

Considering its feasibility and beneficial properties, the steps of our research work could be incorporated into method developments aiming to establish novel diagnostic tools potentially applicable in clinical practice. Our isolation protocol has several advantages, as it does not require expensive equipment or highly trained professionals, and the entire procedure (along with characterization) is performed in about 4 hours.

Although some isolation methods, such as size exclusion chromatography and precipitation, can be performed more quickly, isolation via differential centrifugation results in fewer particles in the size range potentially LPs, lower intensity of LP markers, higher 61–150 nm EVs to 0–60 nm EVs ratio, and higher intensity of EV markers [78]. Raman spectroscopy provides a comprehensive analysis of the circulating tumor-related molecular content. Besides, Raman spectroscopy has additional advantages, such as operator safety, elimination of disposables and analysis waste, fast analytical response of less than 2 min, reduction of the risk of errors because no intrusion or dilution are needed, and negligible maintenance costs. By employing the appropriate preprocessing steps, classification requires reduced computational time and capacity. Moreover, SVM classification based on Raman spectra is suitable to support the proper assessment of even complex biological samples, despite their high degree of variance. This approach may also support decision-making in challenging clinical cases, such as distinguishing between primary brain tumors and other metastatic brain malignancies.

Besides its advantages, our approach also has limiting factors. LPs and protein aggregates in the same size range of sEVs may co-isolate during the differential centrifugation. Accordingly, it is recommended to refer to the isolates as "particle profile with sEVs", "sEV-sized particles", or "sEV-enriched isolates". Because of this heterogeneity, it is also not evident whether the Raman-based classification differentiates the tumor-specific molecular information concentrated in the circulating sEVs or different type of particles. Isolation purity could be improved by combining different isolation methods and examining plasma instead of serum [78].

Isolates from serum may be enriched not only with tumor tissue-derived sEVs but also with EVs released by red blood cells, platelets and immune cells. As it is not revealed whether sEVs from other sources are analytical noise or carriers of relevant information, it might be worthwhile to distinguish tumor tissue-derived sEVs based on surface markers and to perform Raman spectroscopic analyses only on them in the future [76].

In conclusion, our results provide a proof of principle for a novel detection technology that might be utilized to develop a relatively easy-to-execute and appropriate method, which could have the potential to support and simplify the diagnosis and monitoring of CNS tumors in the future. However, clinical applicability definitely requires further development.

4. Materials and Methods

4.1. Patients

Blood samples of 138 patients treated at the Department of Neurosurgery at the University of Debrecen were analyzed. Samples were obtained from patients with GBM, BM, M. Patients with spinal disc herniation (a non-cancerous CNS disease) served as control CTRL (Table 1).

Table 1. Patient cohort.

Patient Groups	No. of Patients	Age (years)			Sex	
		Range	Mean	Median	Male (%)	Female (%)
CTRL	36	20–81	53.6	54	16 (44.4)	20 (55.6)
GBM	46	33–82	64.3	66	28 (60.9)	18 (39.1)
BM	28	42–82	63.5	62.6	18 (64.3)	10 (35.7)
M	28	30–79	58.6	60	5 (17.9)	23 (82.1)

Each patient signed an informed consent form. The study was conducted in accordance with the Declaration of Helsinki, and ethical approval was obtained from two independent bodies (51450-2/2015/EKU (0411/15), Medical Research Council, Scientific and Research Ethics Committee, Budapest, October 30, 2015 and 121/2019-SZTE, University of Szeged, Human Investigation Review Board, Albert Szent-Györgyi Clinical Centre, Szeged, 19 July 2019)

4.2. Preparation of Serum Samples, sEV Isolation and Characterization

Preparation of serum samples was described in our previously published article [20]. Briefly, after 1 h of blood clotting at room temperature, sEV isolation from serum samples was performed via differential centrifugation (20 min at 3000× *g*, 10 °C; 30 min at 10,000× *g*, 4 °C; 70 min at 100,000× *g*, 4 °C). After the last centrifugation step, the pellet was resuspended in Dulbecco's phosphate-buffered saline (DPBS) and was stored at −80 °C until further processing.

To characterize sEVs, we followed the main suggestions and requirements included in the guideline 'Minimal Information for Studies of Extracellular Vesicles 2018' (MISEV 2018) [17].

sEVs were diluted in particle-free DPBS and analyzed using a NanoSight NS300 instrument with 532 nm laser (Malvern Panalytical Ltd., Malvern, UK). Six videos of 60 s were recorded for each sample under constant settings (Camera level: 15; Threshold: 4, 25 °C; 60–80 particles/frame) and analyzed to obtain data on size distribution and particle concentration.

Classical EV markers were presented by Western blot analyses using NuPAGE reagents and an XCell SureLock Mini-Cell System (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's protocols. For detection of the CD81, Alix and Calnexin markers, we used rabbit anti-human CD81 (1:1000, Sigma-Aldrich, St. Louis, MO, USA), rabbit anti-human Alix (1:1000, Sigma-Aldrich, St. Louis, MO, USA) and rabbit anti-human Calnexin (1:10,000), Sigma-Aldrich, St. Louis, MO, USA) primary antibody and HRP-conjugated anti-rabbit IgG (1:1000, R&D Systems, Minneapolis, MN, USA) secondary antibody. THP-1 cell line (ATCC, Teddington, UK) lysate was used for positive control for Calnexin.

In order to examine sEV morphology, TEM analysis was performed using a Tecnai G2 20 X-Twin type instrument (FEI, Hillsboro, OR, USA), operating at an acceleration voltage of 200 kV. For TEM measurements, the samples were dropped on a grid (carbon film with 200 Mesh copper grids (CF200-Cu, Electron Microscopy Sciences, Hatfield, PA, USA) and dried without staining or other fixation procedure.

4.3. Raman Spectroscopy

Raman characterization of sEVs was carried out with a Senterra II microscope (Bruker) in backscattering configuration. The samples were centrifuged, drop-casted on a calcium fluoride substrate and air-dried at room temperature before the analysis. All the samples were analyzed using the same configuration parameters based on preliminary studies: nominal laser power 12.5 mW, integration time 30 s (2 coadditions), interferometer resolution 1.5 cm^{-1}, excitation wavelength 532 nm. The spectra from all the samples were collected by using a 50× optical objective (Olympus). The described optical setup produces a laser spot of approx. 15 µm, which is the sampling area of the Raman spectra, and it is much smaller than the average size of the air-dried sample of approx. 4 mm, thus the Raman microscope operator can finely tune the position of the sampling spot and avoid duplication. The spectra were baseline-corrected before being averaged (5 spectra per sample) using the OPUS software available with the Bruker equipment. Spectral range between 801 cm^{-1} and 3100.5 cm^{-1} was used for further analyses (Table S2).

4.4. Data Adjustment

Row normalization of baseline-corrected data was performed using the SNV method. SNV transformed the mean to 0 and standard deviation to 1, making all spectra comparable in terms of intensity.

PCA with unit variance scaling was applied on the SNV-normalized spectra [79]. PCA served to reduce the dimensions of multivariate data by transforming the original variables (wavenumbers) into a smaller number of new variables, i.e., the PCs.

Data adjustment was performed using the Orange 3.27.0 software (Ljubljana, Slovenia).

4.5. Classification

To develop and test a classification algorithm, the spectra for the samples from the four patient groups were compared pairwise (each patient group was compared to the control, and BM vs. GBM was compared). Sample classification was carried out using the linear SVM algorithm, yielding classification models for each paired group. First, the data were randomly split into train and test sets in a ratio of 90:10. Using the train set, SVM attempts were executed to find a hyperplane that can separate the compared groups in the PCA-transformed space. The process yielded a trained SVM model. Then, the trained SVM model ordered group-membership scores (from 0 to 1) to the test samples based on their positions and distances from the separating hyperplane. In practice, the decisions were made based on the location of the test samples from the plane, which is expressed by their group-membership scores. To make predictions about the test samples, a minimum threshold for the group-membership score was determined. Test samples with scores above this threshold were classified into the target group of interest. In each case, the train–test split was repeated ten times.

Classification efficacy was assessed by sensitivity (proportion of correctly identified positive samples), specificity (proportion of correctly identified negative samples), and by the AUC value obtained from the ROC analysis [66]. Classification and efficacy evaluations were performed using the Orange 3.27.0 and GraphPad Prism 8.4.3 (San Diego, CA, USA) software packages.

4.6. Determining the Spectral Differences

The correlation between the obtained PCs and the different groups was determined by the FreeViz method [80]. Briefly, the FreeViz method displays multivariate data in a 2-dimensional scatterplot to separate samples from different patient groups. In the FreeViz plots, the samples and PCs are represented with dots and vectors, respectively (Figure 4). Since FreeViz optimized the display concerning the patient groups, the PCs that played a more important role in classification generally had longer vectors. Directions of the PC vectors were also revealing. When a region in the graph was mainly populated by samples of a certain group, the PC vectors in that direction could be regarded as good indicators of this group membership. The more a PC vector was approaching perpendicularity relative to the line separating the groups, the more useful it was for distinguishing them. Between-group statistical differences in PCs were analyzed using Welch's *t*-test. Regarding that PCs are the linear combination of the original variables (wavenumbers), it is possible to determine the wavenumbers that have the largest contribution to a given PC.

Values of $p < 0.05$ were considered significant. FreeViz was performed using the Orange 3.78.0 software [81].

5. Conclusions

Our study aimed to classify serum-derived sEVs from four patient groups based on their Raman spectral signatures. To the best of our knowledge, we are the first group to investigate the potential role of serum-derived sEVs in the diagnosis of CNS tumors using Raman spectroscopy. Based on various metrics, the classification efficiency proved to be excellent. In conclusion, our results support that Raman spectroscopic analysis of circulating sEV-enriched isolates is a promising liquid-biopsy-based method that could be further developed in order to be applicable in the diagnosis of CNS tumors. Our easy-to-perform analysis offers a novel detection technology that might be utilized in method developments aiming to simplify the diagnosis and monitoring of CNS tumors, and thus it might have the potential to be integrated into clinical practice in the future.

Supplementary Materials: The following are available online at https://www.mdpi.com/2072-6694/13/6/1407/s1. Figure S1: Original images of the Western blot analysis; Figure S2: PCA score plots of the selected PCs; Table S1: Tabular form of the discriminative spectral differences; Table S2: The baseline-corrected Raman spectroscopic data.

Author Contributions: Conceptualization, K.B.; methodology, M.B., G.D., J.G.-P. and K.B.; validation, K.K., A.J. and S.K.; formal analysis, M.B. and G.D.; investigation, M.B., G.D., J.G.-P., M.H. and E.G.-S.; resources, A.K., P.H. and T.B.; writing—original draft preparation, M.B. and G.D.; writing—review and editing, M.B., G.D., K.B., M.H. and E.G.-S.; visualization, M.B.; supervision, K.B., Z.K. and A.K.; project administration, M.B. and K.B.; funding acquisition, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the following research grants: GINOP-2.3.2-15-2016-00015 (K.B.); GINOP-2.2.1-15-2017-00052 (K.B.), 2017-1.2.1-NKP-2017-00002 "National Brain Research Program NAP 2.0" (A.K.), ÚNKP-19-4-SZTE-63 (K.B.), János Bolyai Research Scholarship of the Hungarian Academic of Sciences (K.B), ÚNKP-20-3-New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund (M.B.), EFOP 3.6.3-VEKOP-16-2017-00009 (M.B.).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and ethical approval was obtained from two independent bodies (51450-2/2015/EKU (0411/15), Medical Research Council, Scientific and Research Ethics Committee, Budapest, 30 October 2015 and 121/2019-SZTE, University of Szeged, Human Investigation Review Board, Albert Szent-Györgyi Clinical Centre, Szeged, 19 July 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All datasets generated during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank Lilla Pintér for her technical assistance. The authors thank Dóra Bokor, for proofreading the manuscript. The authors thank László Janovák and Ágota Imre-Deák for the transmission electron microscopic analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yáñez-Mó, M.; Siljander, P.R.-M.; Andreu, Z.; Bedina Zavec, A.; Borràs, F.E.; Buzas, E.I.; Buzas, K.; Casal, E.; Cappello, F.; Carvalho, J.; et al. Biological Properties of Extracellular Vesicles and Their Physiological Functions. *J. Extracell. Vesicles* **2015**, *4*, 27066. [CrossRef] [PubMed]
2. Takeuchi, T.; Suzuki, M.; Fujikake, N.; Popiel, H.A.; Kikuchi, H.; Futaki, S.; Wada, K.; Nagai, Y. Intercellular Chaperone Transmission via Exosomes Contributes to Maintenance of Protein Homeostasis at the Organismal Level. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E2497–E2506. [CrossRef]
3. Harmati, M.; Gyukity-Sebestyen, E.; Dobra, G.; Janovak, L.; Dekany, I.; Saydam, O.; Hunyadi-Gulyas, E.; Nagy, I.; Farkas, A.; Pankotai, T.; et al. Small Extracellular Vesicles Convey the Stress-Induced Adaptive Responses of Melanoma Cells. *Sci. Rep.* **2019**, *9*, 15329. [CrossRef]
4. Desdín-Micó, G.; Mittelbrunn, M. Role of Exosomes in the Protection of Cellular Homeostasis. *Cell Adhes. Migr.* **2017**, *11*, 127–134. [CrossRef] [PubMed]
5. Mathivanan, S.; Simpson, R.J. ExoCarta: A Compendium of Exosomal Proteins and RNA. *Proteomics* **2009**, *9*, 4997–5000. [CrossRef] [PubMed]
6. Schorey, J.S.; Bhatnagar, S. Exosome Function: From Tumor Immunology to Pathogen Biology. *Traffic* **2008**, *9*, 871–881. [CrossRef] [PubMed]
7. Nogués, L.; Benito-Martin, A.; Hergueta-Redondo, M.; Peinado, H. The Influence of Tumour-Derived Extracellular Vesicles on Local and Distal Metastatic Dissemination. *Mol. Asp. Med.* **2018**, *60*, 15–26. [CrossRef]
8. Hoshino, A.; Costa-Silva, B.; Shen, T.-L.; Rodrigues, G.; Hashimoto, A.; Tesic Mark, M.; Molina, H.; Kohsaka, S.; Di Giannatale, A.; Ceder, S.; et al. Tumour Exosome Integrins Determine Organotropic Metastasis. *Nature* **2015**, *527*, 329–335. [CrossRef]
9. Costa-Silva, B.; Aiello, N.M.; Ocean, A.J.; Singh, S.; Zhang, H.; Thakur, B.K.; Becker, A.; Hoshino, A.; Mark, M.T.; Molina, H.; et al. Pancreatic Cancer Exosomes Initiate Pre-Metastatic Niche Formation in the Liver. *Nat. Cell Biol.* **2015**, *17*, 816–826. [CrossRef]
10. Liu, Y.; Gu, Y.; Han, Y.; Zhang, Q.; Jiang, Z.; Zhang, X.; Huang, B.; Xu, X.; Zheng, J.; Cao, X. Tumor Exosomal RNAs Promote Lung Pre-Metastatic Niche Formation by Activating Alveolar Epithelial TLR3 to Recruit Neutrophils. *Cancer Cell* **2016**, *30*, 243–256. [CrossRef]
11. Zeng, Z.; Li, Y.; Pan, Y.; Lan, X.; Song, F.; Sun, J.; Zhou, K.; Liu, X.; Ren, X.; Wang, F.; et al. Cancer-Derived Exosomal MiR-25-3p Promotes Pre-Metastatic Niche Formation by Inducing Vascular Permeability and Angiogenesis. *Nat. Commun.* **2018**, *9*, 5395. [CrossRef] [PubMed]
12. Feng, W.; Dean, D.C.; Hornicek, F.J.; Shi, H.; Duan, Z. Exosomes Promote Pre-Metastatic Niche Formation in Ovarian Cancer. *Mol. Cancer* **2019**, *18*, 124. [CrossRef] [PubMed]
13. Chen, G.; Huang, A.C.; Zhang, W.; Zhang, G.; Wu, M.; Xu, W.; Yu, Z.; Yang, J.; Wang, B.; Sun, H.; et al. Exosomal PD-L1 Contributes to Immunosuppression and Is Associated with Anti-PD-1 Response. *Nature* **2018**, *560*, 382–386. [CrossRef]
14. Lener, T.; Gimona, M.; Aigner, L.; Börger, V.; Buzas, E.; Camussi, G.; Chaput, N.; Chatterjee, D.; Court, F.A.; del Portillo, H.A.; et al. Applying Extracellular Vesicles Based Therapeutics in Clinical Trials—An ISEV Position Paper. *J. Extracell. Vesicles* **2015**, *4*, 30087. [CrossRef]
15. Ma, C.; Jiang, F.; Ma, Y.; Wang, J.; Li, H.; Zhang, J. Isolation and Detection Technologies of Extracellular Vesicles and Application on Cancer Diagnostic. *Dose Response* **2019**, *17*, 1559325819891004. [CrossRef]
16. Sheridan, C. Exosome Cancer Diagnostic Reaches Market. *Nat. Biotechnol.* **2016**, *34*, 359–360. [CrossRef] [PubMed]
17. Théry, C.; Witwer, K.W.; Aikawa, E.; Alcaraz, M.J.; Anderson, J.D.; Andriantsitohaina, R.; Antoniou, A.; Arab, T.; Archer, F.; Atkin-Smith, G.K.; et al. Minimal Information for Studies of Extracellular Vesicles 2018 (MISEV2018): A Position Statement of the International Society for Extracellular Vesicles and Update of the MISEV2014 Guidelines. *J. Extracell. Vesicles* **2018**, *7*, 1535750. [CrossRef]
18. Colombo, M.; Raposo, G.; Théry, C. Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annu. Rev. Cell Dev. Biol.* **2014**, *30*, 255–289. [CrossRef]
19. Ruhen, O.; Meehan, K. Tumor-Derived Extracellular Vesicles as a Novel Source of Protein Biomarkers for Cancer Diagnosis and Monitoring. *Proteomics* **2019**, *19*, 1800155. [CrossRef]
20. Dobra, G.; Bukva, M.; Szabo, Z.; Bruszel, B.; Harmati, M.; Gyukity-Sebestyen, E.; Jenei, A.; Szucs, M.; Horvath, P.; Biro, T.; et al. Small Extracellular Vesicles Isolated from Serum May Serve as Signal-Enhancers for the Monitoring of CNS Tumors. *IJMS* **2020**, *21*, 5359. [CrossRef]
21. Choy, C.; Jandial, R. Breast Cancer Exosomes Breach the Blood-Brain Barrier. *Neurosurgery* **2016**, *78*, N10–N11. [CrossRef]

22. García-Romero, N.; Carrión-Navarro, J.; Esteban-Rubio, S.; Lázaro-Ibáñez, E.; Peris-Celda, M.; Alonso, M.M.; Guzmán-De-Villoria, J.; Fernández-Carballal, C.; de Mendivil, A.O.; García-Duque, S.; et al. DNA Sequences within Glioma-Derived Extracellular Vesicles Can Cross the Intact Blood-Brain Barrier and Be Detected in Peripheral Blood of Patients. *Oncotarget* **2017**, *8*, 1416–1428. [CrossRef]
23. Scavo, M.P.; Depalo, N.; Tutino, V.; De Nunzio, V.; Ingrosso, C.; Rizzi, F.; Notarnicola, M.; Curri, M.L.; Giannelli, G. Exosomes for Diagnosis and Therapy in Gastrointestinal Cancers. *Int. J. Mol. Sci.* **2020**, *21*, 367. [CrossRef] [PubMed]
24. Melo, S.A.; Luecke, L.B.; Kahlert, C.; Fernandez, A.F.; Gammon, S.T.; Kaye, J.; LeBleu, V.S.; Mittendorf, E.A.; Weitz, J.; Rahbari, N.; et al. Glypican-1 Identifies Cancer Exosomes and Detects Early Pancreatic Cancer. *Nature* **2015**, *523*, 177–182. [CrossRef]
25. Aldape, K.; Brindle, K.M.; Chesler, L.; Chopra, R.; Gajjar, A.; Gilbert, M.R.; Gottardo, N.; Gutmann, D.H.; Hargrave, D.; Holland, E.C.; et al. Challenges to Curing Primary Brain Tumours. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 509–520. [CrossRef]
26. Shankar, G.M.; Balaj, L.; Stott, S.L.; Nahed, B.; Carter, B.S. Liquid Biopsy for Brain Tumors. *Expert Rev. Mol. Diagn.* **2017**, *17*, 943–947. [CrossRef]
27. Garden, G.A.; Campbell, B.M. Glial Biomarkers in Human Central Nervous System Disease: Glial Biomarkers in Human CNS Disease. *Glia* **2016**, *64*, 1755–1771. [CrossRef] [PubMed]
28. Neska-Matuszewska, M.; Bladowska, J.; Sąsiadek, M.; Zimny, A. Differentiation of Glioblastoma Multiforme, Metastases and Primary Central Nervous System Lymphomas Using Multiparametric Perfusion and Diffusion MR Imaging of a Tumor Core and a Peritumoral Zone-Searching for a Practical Approach. *PLoS ONE* **2018**, *13*, e0191341. [CrossRef] [PubMed]
29. Pope, W.B.; Brandal, G. Conventional and Advanced Magnetic Resonance Imaging in Patients with High-Grade Glioma. *Q. J. Nucl. Med. Mol. Imaging* **2018**, *62*, 239–253. [CrossRef] [PubMed]
30. Peca, C.; Pacelli, R.; Elefante, A.; Del Basso De Caro, M.L.; Vergara, P.; Mariniello, G.; Giamundo, A.; Maiuri, F. Early Clinical and Neuroradiological Worsening after Radiotherapy and Concomitant Temozolomide in Patients with Glioblastoma: Tumour Progression or Radionecrosis? *Clin. Neurol. Neurosurg.* **2009**, *111*, 331–334. [CrossRef]
31. Saenz-Antoñanzas, A.; Auzmendi-Iriarte, J.; Carrasco-Garcia, E.; Moreno-Cugnon, L.; Ruiz, I.; Villanua, J.; Egaña, L.; Otaegui, D.; Samprón, N.; Matheu, A. Liquid Biopsy in Glioblastoma: Opportunities, Applications and Challenges. *Cancers* **2019**, *11*, 950. [CrossRef]
32. Patel, A.P.; Tirosh, I.; Trombetta, J.J.; Shalek, A.K.; Gillespie, S.M.; Wakimoto, H.; Cahill, D.P.; Nahed, B.V.; Curry, W.T.; Martuza, R.L.; et al. Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma. *Science* **2014**, *344*, 1396–1401. [CrossRef] [PubMed]
33. Best, M.G.; Sol, N.; Zijl, S.; Reijneveld, J.C.; Wesseling, P.; Wurdinger, T. Liquid Biopsies in Patients with Diffuse Glioma. *Acta Neuropathol.* **2015**, *129*, 849–865. [CrossRef] [PubMed]
34. Good, D.M.; Thongboonkerd, V.; Novak, J.; Bascands, J.-L.; Schanstra, J.P.; Coon, J.J.; Dominiczak, A.; Mischak, H. Body Fluid Proteomics for Biomarker Discovery: Lessons from the Past Hold the Key to Success in the Future. *J. Proteome Res.* **2007**, *6*, 4549–4555. [CrossRef]
35. Marrugo-Ramírez, J.; Mir, M.; Samitier, J. Blood-Based Cancer Biomarkers in Liquid Biopsy: A Promising Non-Invasive Alternative to Tissue Biopsy. *Int. J. Mol. Sci.* **2018**, *19*, 2877. [CrossRef]
36. Miyauchi, E.; Furuta, T.; Ohtsuki, S.; Tachikawa, M.; Uchida, Y.; Sabit, H.; Obuchi, W.; Baba, T.; Watanabe, M.; Terasaki, T.; et al. Identification of Blood Biomarkers in Glioblastoma by SWATH Mass Spectrometry and Quantitative Targeted Absolute Proteomics. *PLoS ONE* **2018**, *13*, e0193799. [CrossRef] [PubMed]
37. Lin, B.; White, J.T.; Wu, J.; Lele, S.; Old, L.J.; Hood, L.; Odunsi, K. Deep Depletion of Abundant Serum Proteins Reveals Low-Abundant Proteins as Potential Biomarkers for Human Ovarian Cancer. *Prot. Clin. Appl.* **2009**, *3*, 853–861. [CrossRef] [PubMed]
38. Cagney, D.N.; Sul, J.; Huang, R.Y.; Ligon, K.L.; Wen, P.Y.; Alexander, B.M. The FDA NIH Biomarkers, EndpointS, and Other Tools (BEST) Resource in Neuro-Oncology. *Neuro Oncol.* **2018**, *20*, 1162–1172. [CrossRef] [PubMed]
39. Gollapalli, K.; Ray, S.; Srivastava, R.; Renu, D.; Singh, P.; Dhali, S.; Bajpai Dikshit, J.; Srikanth, R.; Moiyadi, A.; Srivastava, S. Investigation of Serum Proteome Alterations in Human Glioblastoma Multiforme. *Proteomics* **2012**, *12*, 2378–2390. [CrossRef] [PubMed]
40. Figueroa, J.M.; Carter, B.S. Detection of Glioblastoma in Biofluids. *J. Neurosurg.* **2018**, *129*, 334–340. [CrossRef] [PubMed]
41. Choi, D.; Montermini, L.; Kim, D.-K.; Meehan, B.; Roth, F.P.; Rak, J. The Impact of Oncogenic EGFRvIII on the Proteome of Extracellular Vesicles Released from Glioblastoma Cells. *Mol. Cell Proteom.* **2018**, *17*, 1948–1964. [CrossRef] [PubMed]
42. Pence, I.; Mahadevan-Jansen, A. Clinical Instrumentation and Applications of Raman Spectroscopy. *Chem. Soc. Rev.* **2016**, *45*, 1958–1979. [CrossRef] [PubMed]
43. Gualerzi, A.; Niada, S.; Giannasi, C.; Picciolini, S.; Morasso, C.; Vanna, R.; Rossella, V.; Masserini, M.; Bedoni, M.; Ciceri, F.; et al. Raman Spectroscopy Uncovers Biochemical Tissue-Related Features of Extracellular Vesicles from Mesenchymal Stromal Cells. *Sci. Rep.* **2017**, *7*, 9820. [CrossRef] [PubMed]
44. Maisano, D.; Mimmi, S.; Russo, R.; Fioravanti, A.; Fiume, G.; Vecchio, E.; Nisticò, N.; Quinto, I.; Iaccino, E. Uncovering the Exosomes Diversity: A Window of Opportunity for Tumor Progression Monitoring. *Pharmaceuticals* **2020**, *13*, 180. [CrossRef]
45. Harris, A.T.; Lungari, A.; Needham, C.J.; Smith, S.L.; Lones, M.A.; Fisher, S.E.; Yang, X.B.; Cooper, N.; Kirkham, J.; Smith, D.A.; et al. Potential for Raman Spectroscopy to Provide Cancer Screening Using a Peripheral Blood Sample. *Head Neck Oncol.* **2009**, *1*, 34. [CrossRef]

46. Pichardo-Molina, J.L.; Frausto-Reyes, C.; Barbosa-García, O.; Huerta-Franco, R.; González-Trujillo, J.L.; Ramírez-Alvarado, C.A.; Gutiérrez-Juárez, G.; Medina-Gutiérrez, C. Raman Spectroscopy and Multivariate Analysis of Serum Samples from Breast Cancer Patients. *Lasers Med. Sci.* **2007**, *22*, 229–236. [CrossRef]
47. Mehta, K.; Atak, A.; Sahu, A.; Srivastava, S.; Krishna C, M. An Early Investigative Serum Raman Spectroscopy Study of Meningioma. *Analyst* **2018**, *143*, 1916–1923. [CrossRef] [PubMed]
48. Park, J.; Hwang, M.; Choi, B.; Jeong, H.; Jung, J.; Kim, H.K.; Hong, S.; Park, J.; Choi, Y. Exosome Classification by Pattern Analysis of Surface-Enhanced Raman Spectroscopy Data for Lung Cancer Diagnosis. *Anal. Chem.* **2017**, *89*, 6695–6701. [CrossRef]
49. Carmicheal, J.; Hayashi, C.; Huang, X.; Liu, L.; Lu, Y.; Krasnoslobodtsev, A.; Lushnikov, A.; Kshirsagar, P.G.; Patel, A.; Jain, M.; et al. Label-Free Characterization of Exosome via Surface Enhanced Raman Spectroscopy for the Early Detection of Pancreatic Cancer. *Nanomed. Nanotechnol. Biol. Med.* **2019**, *16*, 88–96. [CrossRef]
50. Ostrom, Q.T.; Gittleman, H.; Truitt, G.; Boscia, A.; Kruchko, C.; Barnholtz-Sloan, J.S. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011–2015. *Neuro Oncol.* **2018**, *20*, iv1–iv86. [CrossRef]
51. Fox, B.D.; Cheung, V.J.; Patel, A.J.; Suki, D.; Rao, G. Epidemiology of Metastatic Brain Tumors. *Neurosurg. Clin. N. Am.* **2011**, *22*, 1–6. [CrossRef] [PubMed]
52. Zhi, F.; Shao, N.; Li, B.; Xue, L.; Deng, D.; Xu, Y.; Lan, Q.; Peng, Y.; Yang, Y. A Serum 6-MiRNA Panel as a Novel Non-Invasive Biomarker for Meningioma. *Sci. Rep.* **2016**, *6*, 32067. [CrossRef]
53. Taverna, S.; Giallombardo, M.; Gil-Bazo, I.; Carreca, A.P.; Castiglia, M.; Chacártegui, J.; Araujo, A.; Alessandro, R.; Pauwels, P.; Peeters, M.; et al. Exosomes Isolation and Characterization in Serum Is Feasible in Non-Small Cell Lung Cancer Patients: Critical Analysis of Evidence and Potential Role in Clinical Practice. *Oncotarget* **2016**, *7*, 28748–28760. [CrossRef] [PubMed]
54. Osti, D.; Del Bene, M.; Rappa, G.; Santos, M.; Matafora, V.; Richichi, C.; Faletti, S.; Beznoussenko, G.V.; Mironov, A.; Bachi, A.; et al. Clinical Significance of Extracellular Vesicles in Plasma from Glioblastoma Patients. *Clin. Cancer Res.* **2019**, *25*, 266–276. [CrossRef]
55. Lázaro-Ibáñez, E.; Sanz-Garcia, A.; Visakorpi, T.; Escobedo-Lucea, C.; Siljander, P.; Ayuso-Sacido, Á.; Yliperttula, M. Different GDNA Content in the Subpopulations of Prostate Cancer Extracellular Vesicles: Apoptotic Bodies, Microvesicles, and Exosomes. *Prostate* **2014**, *74*, 1379–1390. [CrossRef]
56. König, L.; Kasimir-Bauer, S.; Bittner, A.-K.; Hoffmann, O.; Wagner, B.; Santos Manvailer, L.F.; Kimmig, R.; Horn, P.A.; Rebmann, V. Elevated Levels of Extracellular Vesicles are Associated with Therapy Failure and Disease Progression in Breast Cancer Patients Undergoing Neoadjuvant Chemotherapy. *Oncoimmunology* **2018**, *7*, e1376153. [CrossRef]
57. Gercel-Taylor, C.; Atay, S.; Tullis, R.H.; Kesimer, M.; Taylor, D.D. Nanoparticle Analysis of Circulating Cell-Derived Vesicles in Ovarian Cancer Patients. *Anal. Biochem.* **2012**, *428*, 44–53. [CrossRef]
58. Liu, M.-L.; Werth, V.P.; Williams, K.J. Blood Plasma versus Serum: Which Is Right for Sampling Circulating Membrane Microvesicles in Human Subjects? *Ann. Rheum. Dis.* **2019**, *79*, e73. [CrossRef]
59. Ji, Q.; Ji, Y.; Peng, J.; Zhou, X.; Chen, X.; Zhao, H.; Xu, T.; Chen, L.; Xu, Y. Increased Brain-Specific MiR-9 and MiR-124 in the Serum Exosomes of Acute Ischemic Stroke Patients. *PLoS ONE* **2016**, *11*, e0163645. [CrossRef] [PubMed]
60. Cunha, C.; Silva, A.J.; Pereira, P.; Vaz, R.; Gonçalves, R.M.; Barbosa, M.A. The Inflammatory Response in the Regression of Lumbar Disc Herniation. *Arthritis Res. Ther.* **2018**, *20*, 251. [CrossRef] [PubMed]
61. Gardiner, C.; Harrison, P.; Belting, M.; Böing, A.; Campello, E.; Carter, B.S.; Collier, M.E.; Coumans, F.; Ettelaie, C.; van Es, N.; et al. Extracellular Vesicles, Tissue Factor, Cancer and Thrombosis—Discussion Themes of the ISEV 2014 Educational Day. *J. Extracell. Vesicles* **2015**, *4*, 26901. [CrossRef] [PubMed]
62. Anderson, K.S.; LaBaer, J. The Sentinel Within: Exploiting the Immune System for Cancer Biomarkers †. *J. Proteome Res.* **2005**, *4*, 1123–1133. [CrossRef] [PubMed]
63. Wen, C.; Seeger, R.C.; Fabbri, M.; Wang, L.; Wayne, A.S.; Jong, A.Y. Biological Roles and Potential Applications of Immune Cell-Derived Extracellular Vesicles. *J. Extracell. Vesicles* **2017**, *6*, 1400370. [CrossRef]
64. Salem, N.; Hussein, S. Data Dimensional Reduction and Principal Components Analysis. *Procedia Comput. Sci.* **2019**, *163*, 292–299. [CrossRef]
65. Lyng, F.M.; Traynor, D.; Nguyen, T.N.Q.; Meade, A.D.; Rakib, F.; Al-Saady, R.; Goormaghtigh, E.; Al-Saad, K.; Ali, M.H. Discrimination of Breast Cancer from Benign Tumours Using Raman Spectroscopy. *PLoS ONE* **2019**, *14*, e0212376. [CrossRef]
66. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627–635.
67. Mandrekar, J.N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [CrossRef]
68. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
69. Zheng, C.; Qing, S.; Wang, J.; Lü, G.; Li, H.; Lü, X.; Ma, C.; Tang, J.; Yue, X. Diagnosis of Cervical Squamous Cell Carcinoma and Cervical Adenocarcinoma Based on Raman Spectroscopy and Support Vector Machine. *Photodiagn. Photodyn. Ther.* **2019**, *27*, 156–161. [CrossRef]
70. Li, S.; Guo, Z.; Liu, Z. Surface-Enhanced Raman Spectroscopy + Support Vector Machine: A New Noninvasive Method for Prostate Cancer Screening? *Expert Rev. Anticancer Ther.* **2015**, *15*, 5–7. [CrossRef]
71. Socrates, G. *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*, 3rd ed.; John Wiley & Sons Ltd: Chichester, UK, 2010; ISBN 9780470093078.

72. Zhou, W.; Wahl, D.R. Metabolic Abnormalities in Glioblastoma and Metabolic Strategies to Overcome Treatment Resistance. *Cancers* **2019**, *11*, 1231. [CrossRef]
73. Majem, B.; Nadal, E.; Muñoz-Pinedo, C. Exploiting Metabolic Vulnerabilities of Non Small Cell Lung Carcinoma. *Semin. Cell Dev. Biol.* **2020**, *98*, 54–62. [CrossRef]
74. Smolarz, M.; Pietrowska, M.; Matysiak, N.; Mielańczyk, Ł.; Widłak, P. Proteome Profiling of Exosomes Purified from a Small Amount of Human Serum: The Problem of Co-Purified Serum Components. *Proteomes* **2019**, *7*, 18. [CrossRef]
75. Sartori, M.T.; Della Puppa, A.; Ballin, A.; Saggiorato, G.; Bernardi, D.; Padoan, A.; Scienza, R.; d'Avella, D.; Cella, G. Prothrombotic State in Glioblastoma Multiforme: An Evaluation of the Procoagulant Activity of Circulating Microparticles. *J. Neurooncol.* **2011**, *104*, 225–231. [CrossRef]
76. Redzic, J.S.; Ung, T.H.; Graner, M.W. Glioblastoma Extracellular Vesicles: Reservoirs of Potential Biomarkers. *Pharm. Pers. Med.* **2014**, *7*, 65–77. [CrossRef]
77. Enciso-Martinez, A.; Van Der Pol, E.; Hau, C.M.; Nieuwland, R.; Van Leeuwen, T.G.; Terstappen, L.W.M.M.; Otto, C. Label-Free Identification and Chemical Characterisation of Single Extracellular Vesicles and Lipoproteins by Synchronous Rayleigh and Raman Scattering. *J. Extracell. Vesicles* **2020**, *9*, 1730134. [CrossRef]
78. Brennan, K.; Martin, K.; FitzGerald, S.P.; O'Sullivan, J.; Wu, Y.; Blanco, A.; Richardson, C.; Mc Gee, M.M. A Comparison of Methods for the Isolation and Separation of Extracellular Vesicles from Protein and Lipid Particles in Human Serum. *Sci. Rep.* **2020**, *10*, 1039. [CrossRef]
79. Rinnan, Å.; van den Berg, F.; Engelsen, S.B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
80. Demšar, J.; Leban, G.; Zupan, B. FreeViz—An Intelligent Multivariate Visualization Approach to Explorative Analysis of Biomedical Data. *J. Biomed. Inform.* **2007**, *40*, 661–671. [CrossRef]
81. Demšar, J.; Tomaz, C.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; StariC, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn.* **2013**, *14*, 2349–2353.

Cell Communication and Signaling

RESEARCH Open Access

Machine learning-based analysis of cancer cell-derived vesicular proteins revealed significant tumor-specificity and predictive potential of extracellular vesicles for cell invasion and proliferation – A meta-analysis

Matyas Bukva[1,2,3], Gabriella Dobra[1,2,3], Edina Gyukity-Sebestyen[1,3], Timea Boroczky[1,2,3], Marietta Margareta Korsos[1], David G. Meckes Jr.[4], Peter Horvath[3], Krisztina Buzas[1,3] and Maria Harmati[1,3*]

Abstract

Background Although interest in the role of extracellular vesicles (EV) in oncology is growing, not all potential aspects have been investigated. In this meta-analysis, data regarding (i) the EV proteome and (ii) the invasion and proliferation capacity of the NCI-60 tumor cell lines (60 cell lines from nine different tumor types) were analyzed using machine learning methods.

Methods On the basis of the entire proteome or the proteins shared by all EV samples, 60 cell lines were classified into the nine tumor types using multiple logistic regression. Then, utilizing the Least Absolute Shrinkage and Selection Operator, we constructed a discriminative protein panel, upon which the samples were reclassified and pathway analyses were performed. These panels were validated using clinical data ($n = 4{,}665$) from Human Protein Atlas.

Results Classification models based on the entire proteome, shared proteins, and discriminative protein panel were able to distinguish the nine tumor types with 49.15%, 69.10%, and 91.68% accuracy, respectively. Invasion and proliferation capacity of the 60 cell lines were predicted with $R^2 = 0.68$ and $R^2 = 0.62$ ($p < 0.0001$). The results of the Reactome pathway analysis of the discriminative protein panel suggest that the molecular content of EVs might be indicative of tumor-specific biological processes.

Conclusion Integrating in vitro EV proteomic data, cell physiological characteristics, and clinical data of various tumor types illuminates the diagnostic, prognostic, and therapeutic potential of EVs.

Keywords Extracellular vesicles, NCI-60, Invasion, Proliferation, Classification, Prediction, Machine learning

*Correspondence:
Maria Harmati
harmatimarcsi@gmail.com
Full list of author information is available at the end of the article

Background

Cancer growth, progression and metastasis are associated with genomic, proteomic, transcriptomic and metabolomic changes [1]. Omics sciences such as genomics, proteomics, transcriptomics and metabolomics are revolutionizing the understanding of cancer by comparing vast amounts of data with clinical features [2, 3]. Sources of data include in vitro experiments [4], clinical samples [5] and liquid biopsies [6], but nowadays extracellular vesicles (EVs) are of increasing interest due to their role in cell-to-cell communication, as they influence various physiological processes, including tumor-related functions such as immune regulation, cancer cell support, angiogenesis and metastasis [7–9].

Our research, along with others, suggests that EVs have great potential as a source of biomarkers that could advance the current state of cancer diagnosis because they provide a membrane-protected cargo that could reflect cell-specific pathological processes [10–14].

Numerous studies have highlighted the role of EVs in tumorous processes, leading to efforts to include them in liquid biopsy based diagnostic methods [15].

The majority of these studies have demonstrated that the analysis of EVs – due to the tumor-associated molecular pattern carried – can be used to differentiate between tumorous and control samples or to subcategorize tumor types based on their properties (e.g. chemosensitivity) [16–25].

However, there are still a number of unexplored areas regarding the potential utility of EVs. For instance, it is still under exploration whether the molecular composition of EVs can predict the invasion capacity or proliferation rate of the donor cells, or whether they could provide information on tumor-specific signaling pathways or strategies. Furthermore, as most of the studies investigate a limited number of groups, the degree of specificity of the molecular pattern carried by EVs of different tumor types is not fully elucidated.

Comprehensive studies of EVs derived from different tumor types are needed to fully explore their potential use in clinical practice. As a result, in recent years, there has been a rise in research into the proteome of EVs derived from the highly diverse NCI-60 cell line panel compiled by the National Cancer Institute. Using omics approaches to investigate the NCI-60 cell line panel, which contains 60 cell lines from nine tumor types, has significantly contributed to the discovery of potential biomarkers and drug targets, as well as understanding the molecular basis of chemotherapy resistance [26–40].

Beyond the research on the cell lysates, proteomic analysis of EVs of the NCI-60 cell lines revealed that their protein content reflects the molecular composition of the progenitor cell at both the proteomic and transcriptomic levels [41]. EVs were discovered to contain components of the core vesicle machinery, biomarkers already known from tissue, and integrin content that may be tumor stage-specific [41, 42].

Yet as omics and clinical data volumes rise, so do advances in information-processing tools, such as novel machine learning methods and advances in bioinformatics [43].

With this in mind, we hypothesized that we could mine valuable information on the role of EVs in tumor processes by comparing publicly available NCI-60 EV proteomics, cell physiology and clinical data using machine learning and the latest bioinformatics methods.

In our meta-analysis, we created classification models based on the entire proteome identified in the NCI-60 EVs as well as the proteins commonly identified in all samples. Using a selection algorithm, we compiled a panel of the most discriminative proteins from the entire proteome. Thereafter, we conducted enrichment analyses to determine which signal pathways our discriminative proteins are associated with these discriminative proteins. Furthermore, we assembled protein panels capable of estimating the invasion capacity and proliferation rate of donor cells, and validated them with in vivo clinical data.

Materials and methods

Data set used

Proteomic data

We obtained the proteomic data of EVs from the publication of Hurwitz et al. as freely downloadable supplementary material [41]. This data set contains the spectral count and intensity of 6,701 proteins for 60 EV isolates harvested from 60 cell lines (NCI-60) of nine different tumor types. In our study, we used the intensity values for the analyses. Before the analyses, the intensities were logarithmized in order to increase the linearity and reduce the variance. Imputation of missing values was not performed, as the 0 values in the data matrix used do not represent missing values, but the absence of proteins in the EV isolate.

Data on the invasion capacity of NCI-60 cell lines

The invasion phenotype of the 60 cell lines were obtained from the publication of DeLosh et al. as freely downloadable supplementary material [44].

Briefly, DeLosh et al. utilized CIM (cellular invasion/migration)-Plate 16 to determine the invasion capacity of the NCI-60 panel.

The CIM Plate-16 consists of two chambers, one below the other. The chambers are separated by a microporous membrane. Microelectronic sensors are integrated at the bottom of the pores in the lower chamber on the other side of the membrane. The migration of cells from the upper chamber to the lower chamber in response to a chemoattractant leads to their interaction and attachment to the electrical sensors, hence causing an elevation in impedance. The impedance correlates to increasing numbers of migrated cells on the underside of the membrane, and cell index values reflecting impedance changes are automatically and continuously recorded by the Roche xCELLigence Real-Time Cell Analyzer DP instrument. Therefore, cell migration activity can be monitored via the cell index profile.

The invasion phenotype of 60 cell lines was determined by plotting the cell index (reflecting the mass of the cell detected) as a function of analysis time and then calculating the area under the curve (AUC). We used the average AUC for each cell line as published in the original article, but refer to it as invasion capacity for ease of interpretation.

Data on the proliferation of NCI-60 cell lines

Doubling time of NCI-60 cell lines data were obtained from the National Cancer Institute website although [45], to facilitate interpretation, we refer to it as proliferation capacity for ease of interpretation.

Data on RNA expression of the NCI-60 cell lines

Microarray gene expression data was downloaded from the NCBI Gene Expression Omnibus (accession number: GSE32474) [46].

Data on the in situ tissue expression and survival data

In our study, we acquired information from the Human Protein Atlas database regarding the ex vivo tissue expression of specific proteins and the overall survival time (in years) of patients corresponding to the tissue samples [47].

Classification of EV samples

During the classification, we attempted to classify the 60 EV samples into their respective nine tumor types (breast, central nervous system—CNS, colon, kidney, leukemia, lung, melanoma, ovary, prostate).

We applied multiple logistic regression on the proteomic data set for classification purposes.

First, the 60 EV sample was classified based on shared proteins and then on the entire proteome.

After classifying based on the entire proteome, we aimed to identify a discriminant protein panel for the nine tumor types.

The data set was split 50–50%, creating a Train and a Test set. We utilized the Least Absolute Shrinkage and Selection Operator (LASSO) method to score the proteins on the Train set according to their importance in distinguishing the tumor types (this score is the regression coefficients). This value can be negative, positive, or zero, suggesting a negative or positive effect on the probability of classifying into a certain tumor type, or an irrelevant protein.

In LASSO, the so-called cost strength parameter (C), which can vary from 0.001 to 1000, indicates how strict the scoring is (affecting the number of proteins scored as irrelevant/meaningless). In this study, this value was set to 1, which resulted in neither too strong nor too weak scoring, and allowed us to select characteristic proteins for each of the nine tumor types. The optimal value of the parameter C was determined by five-fold cross-validation of the train set and fixed at the point where the highest classification efficiency was measured.

The list of characteristic proteins for the nine tumor types included only proteins with a positive score obtained by LASSO. Classification was again performed on the Test data set based on the proteins selected.

The efficiency of the classification was given by the classification accuracy (number of correctly classified samples divided by the total number of samples). The success of the classification was visualized using confusion matrices.

Orange 3.27.0 [48] software was used to conduct the classification and create figures.

Regression for invasion and proliferation capacity

To predict invasion and proliferation capacity, multiple linear regression with LASSO (with parameter $C=1$) was performed. For regression, LASSO played the same role as in classification.

It should be noted that the approach (CIM Plate-16) used to determine invasiveness of the cell lines has been shown to be applicable only to solid tumors [44], therefore leukemia was not included in the determination of proteins predictive of invasion capacity.

During the procedure, the data was split 50–50%, creating a Train and Test set. On the Train set, LASSO was used to identify proteins that could potentially predict invasion and proliferation capacity. Then, using the Test set, the relationship between the selected proteins and invasion/proliferation capacity was investigated by multiple linear regression.

Value of $p < 0.05$ was considered significant.

The efficiency of the regression was given by the coefficient of determination (R^2).

Orange 3.27.0, GraphPad Prism 8.4.3 (San Diego, CA, USA) were used for multiple linear regression and visualization.

Pathway enrichment analysis

We utilized ShinyGO 0.76.3 for Gene Ontology Enrichment Analysis to determine the biological processes, molecular functions, and cellular components whose proteins are overrepresented in our data set [49]. The ShinyGO parameters were set to default.

Reactome (v82) was employed for simultaneous enrichment analysis of each sample in order to compare the 60 EV samples in terms of their associated signal pathways [50]. The Reactome parameters were set to default.

Value of $p < 0.05$ corrected with the false discovery rate (FDR) method was considered significant.

Hierarchical clustering

Hierarchical clustering based on proteins was performed after row centering and unit variance scaling. Both rows (proteins) and columns (EV samples) were clustered using correlation distance and complete linkage.

Hierarchical clustering based on the Reactome results was performed on raw data, without any adjustment. The rows (pathways) were clustered using correlation distance and complete linkage.

Hierarchical clustering was performed using Morpheus software [51].

T-distributed stochastic neighbor embedding

In order to visualize the proteomic data in a 2-dimensional space, we utilized the t-distributed stochastic neighbor embedding (t-SNE) method.

For t-SNE visualization, we used Orange 3.27.0.

Examining the similarity between the EV proteome and the cellular RNA profile

The similarity of protein and RNA profiles of EV samples and cells for each variable was tested by Spearman's correlation analysis, the results of which were plotted on heatmaps. In addition, the concordance of the two matrices (RNA profile of cells and protein content of EVs) was characterized overall with R_V coefficients introduced by Escoufier [52].

In data analysis, the R_V coefficient is a multivariate generalization of the squared correlation coefficient, depicting the similarity between two matrices of quantitative variables. The R_V coefficient takes values between 0 and 1.

The analysis was performed using the *omicade4* package in the R statistical framework [53].

Survival analysis

The association between tissue expression of certain proteins and survival was determined by Kaplan–Meier analysis with logrank test, using GraphPad Prism 8.4.3. Value of $p < 0.05$ was considered significant.

Results

Machine learning methods revealed tumor-specific protein patterns of EV proteome

Shared proteins of EVs are related to EV biogenesis processes

The proteomic data set of the 60 EV samples contained 6,071 proteins. Intensity was measured for 5,908 proteins, referred to as the entire proteome in this study.

According to Gene Ontology Enrichment Analysis, the entire proteome is significantly associated with biological processes, molecular functions and cellular compartments such as neutrophil-mediated immunity, cell adhesion to the extracellular matrix, secretory vesicles and granules (Additional file 1). The fold enrichment values—which indicates how drastically genes of a certain pathway are overrepresented—ranged between 1.68 and 3.01. This means that we identified at least 1.68 times more proteins from the listed signal pathways as it would have been expected by chance.

Of the 5,908 proteins, 213 were present in all EV samples, referred to as the core proteome. The enrichment analysis of the core proteome showed that the shared proteins are involved in intracellular and EV biogenesis pathways, such as cotranslational protein targeting to membrane, RNA binding and cytosolic ribosomes (Additional file 2). Association of the core proteome with each biological pathway showed higher significance than the entire proteome, which was reflected in the fold enrichment values ranging from 3.78 to 33.12.

Entire proteome of EVs resulted higher classification accuracy of tumor cell lines than core proteome

We first inspected the core proteome for tumor-specific patterns using the logistic regression classification model.

Remarkably, even this small subset of the entire proteome affecting a few biological processes carried enough specific information to distinguish certain tumor types from the others to some extent, such as kidney, lung, leukemia and melanoma (Fig. 1a, c). The classification accuracy of 49.14% significantly outperformed the 11.1% that would have been obtained with random classification.

As expected, a one-way ANOVA analysis revealed that the average intensity of the core proteome depends on tumor type ($p < 0.0001$). However, Pearson's correlation analyses confirmed that this difference could not be caused by differences in EV secretion, EV mean and mode size, or cell size. No significant correlation was identified between any parameter and the average intensity of the core proteome. This suggests that the unique core proteome pattern is not caused by the difference in EV production rate and type of EVs between the nine tumor types, but the different tissue origin.

Using the entire proteome, the distinction between tumor types had become even more defined (Fig. 1b, d). Classification accuracy significantly increased for CNS, colon, leukemia, lung, melanoma, and ovary. The average classification accuracy increased to 69.10% which is 57.99% higher than chance.

The EV proteome could be used to form a discriminative protein panel

In exploring the discriminatory protein panel, we have taken care to ensure that the method does not become overestimated or overfitted. To achieve this, the 60 cell lines were split 50–50%. On one half of the cell lines, the Train set, we applied the LASSO algorithm.

Using the LASSO method, we were able to assign importance scores to each protein of the entire proteome based on their ability to differentiate the nine tumor types in the Train set. The selection algorithm (with parameter $C=1$) resulted in 172 proteins, which were further investigated for hierarchical clustering, classification purposes and Reactome pathway analysis (Additional file 3).

In the hierarchical clustering, the Train and Test sets were analyzed together on the basis of 172 proteins.

Hierarchical clustering using a heatmap revealed that the 172 proteins form a well-defined pattern, enabling the 60 EV samples to form nearly perfectly homogenous clusters, while the Train and Test sets elements are clustered together (Fig. 2a).

This separation is also evident in the t-SNE plots, which depict the various tumor types as distinct groups (Fig. 2b). Again, the elements of the Train and Test sets populated the same areas.

When the samples of the Test set were classified based on the 172 proteins, an average classification efficiency of 91.67% was achieved (Fig. 2c).

For the whole data set (Train + Test), the average efficiency was 96.60%.

Discriminative proteins might uncover tumor-specific pathways

After selecting the proteins, we hypothesized that – given the proteins' large intergroup differences – the biological signaling pathways they affect would also exhibit distinctive patterns. In order to place the 172 selected proteins in a biological context Reactome enrichment analysis was utilized. Only those pathways with $p<0.05$ were considered for hierarchical clustering and heatmap creation (Fig. 3).

The selected 172 proteins are associated with extracellular matrix, nuclear processes, and cell division-related signaling pathways.

Although cancers of the breast and prostate lacked characteristic signaling pathways, the majority of the EV samples clustered according to their tumor type revealing a distinctive signaling pathway pattern.

The collagen matrix, TGF-β receptor, and ERB4 enzyme signaling pathways were identified as common characteristics for both kidney and central nervous system tumors, which clustered together.

Compared to other tumors, leukemia samples exhibit a predominance of nuclear processes associated with histone and chromatin modification.

In general, lung tumors were distinguished by platelet-associated biological processes and integrin-signaling pathways.

Extracellular vesicles carry information on invasion and proliferation capacity

The NCI-60 cell line panel contains not only tumors of different tissue origin, but also tumors with different invasion capacities and different division rates.

Noting that tumor cell lines with low invasion capacity such as BT549 and Hs 578 T (breast) were classified into tumors with high invasion capacity (e.g. CNS) during classification and hierarchical clustering the question arose whether further protein panels predicting invasion and proliferation capacity could be defined.

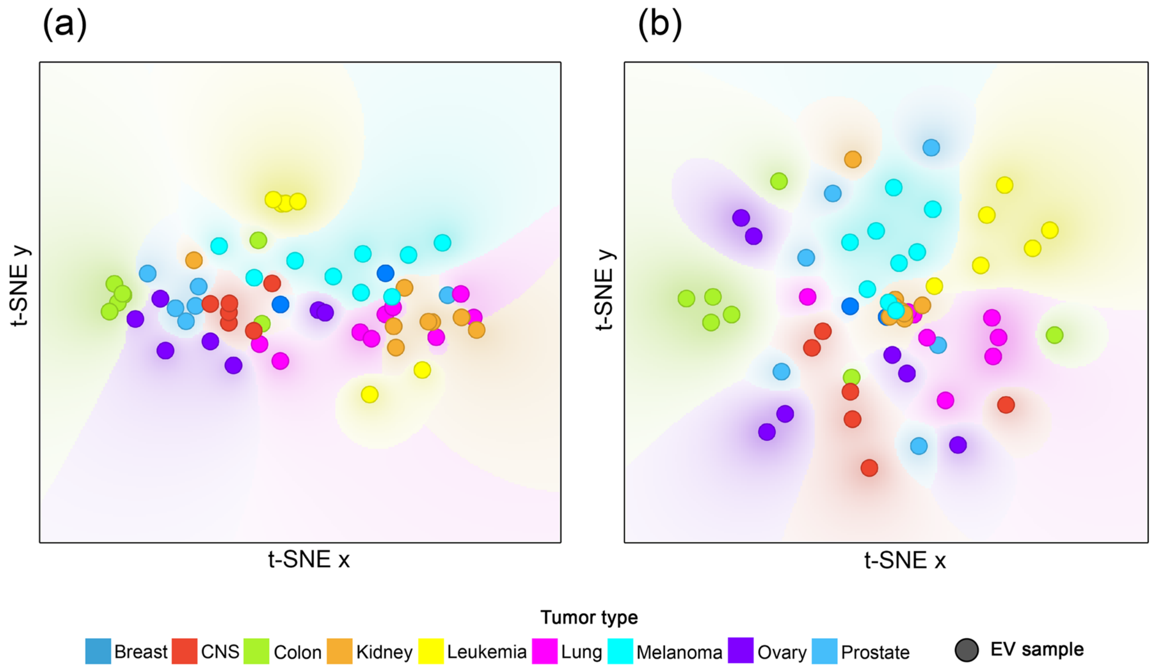
To construct a panel correlated with invasion and proliferation capacity, multiple linear regression with LASSO selection method was utilized.

As in the classification procedure, the data set was split 50–50%. On the Train set, we used LASSO to identify proteins that could be predictive for invasion capacity and proliferation, then validated the findings on the Test set.

The selection resulted in 20 and 15 proteins, which tended to have predictive potential for invasion and

(See figure on next page.)

Fig. 1 Classification efficiency based on the core and entire proteome. **a** t-SNE plot of the core proteome. **b** t-SNE plot of the entire proteome. The dots with different colors represent the 60 individual EV samples belonging to the nine tumor types. The color gradient in the plot indicates the dot density. **c** Confusion matrix of the classification results using the core proteome. **d** Confusion matrix of the classification results using the entire proteome. Each row of the matrices represents the instances in an actual class while each column represents the instances in a predicted class. Diagonally, the percentage of the correct classification is shown in blue. The percentage of errors is indicated in red



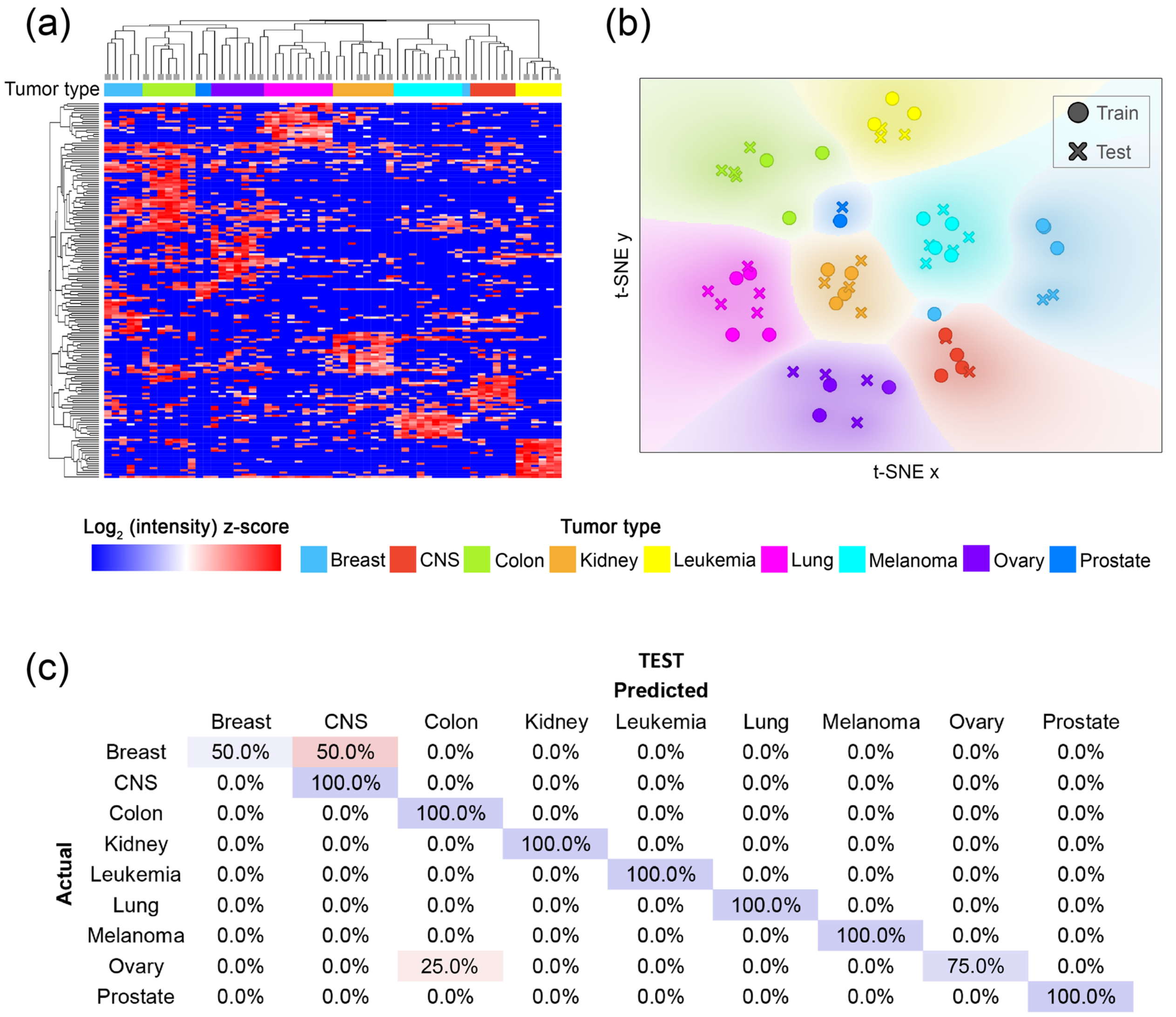
(c)

Actual \ Predicted	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	33.3 %	33.3 %	16.7 %	0.0 %	0.0 %	0.0 %	0.0 %	16.7 %	0.0 %
CNS	16.7 %	50.0 %	0.0 %	16.7 %	0.0 %	16.7 %	0.0 %	0.0 %	0.0 %
Colon	28.6 %	14.3 %	57.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Kidney	0.0 %	0.0 %	0.0 %	87.5 %	0.0 %	12.5 %	0.0 %	0.0 %	0.0 %
Leukemia	0.0 %	0.0 %	0.0 %	16.7 %	66.7 %	0.0 %	16.7 %	0.0 %	0.0 %
Lung	0.0 %	0.0 %	0.0 %	11.1 %	11.1 %	66.7 %	0.0 %	11.1 %	0.0 %
Melanoma	0.0 %	0.0 %	11.1 %	0.0 %	0.0 %	0.0 %	66.7 %	22.2 %	0.0 %
Ovary	28.6 %	0.0 %	14.3 %	14.3 %	0.0 %	0.0 %	28.6 %	14.3 %	0.0 %
Prostate	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %

(d)

Actual \ Predicted	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	33.3 %	33.3 %	16.7 %	0.0 %	0.0 %	0.0 %	0.0 %	16.7 %	0.0 %
CNS	16.7 %	83.3 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Colon	0.0 %	0.0 %	71.4 %	28.6 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Kidney	12.5 %	0.0 %	0.0 %	87.5 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Leukemia	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Lung	0.0 %	0.0 %	11.1 %	0.0 %	0.0 %	88.9 %	0.0 %	0.0 %	0.0 %
Melanoma	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %
Ovary	14.3 %	14.3 %	14.3 %	0.0 %	0.0 %	0.0 %	0.0 %	57.1 %	0.0 %
Prostate	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %

Fig. 1 (See legend on previous page.)



TEST

Predicted

Actual	Breast	CNS	Colon	Kidney	Leukemia	Lung	Melanoma	Ovary	Prostate
Breast	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CNS	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Colon	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Kidney	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Leukemia	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Lung	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Melanoma	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Ovary	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	0.0%	75.0%	0.0%
Prostate	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Fig. 2 Classification efficiency for the selected proteins. **a** Heatmap with hierarchical clustering. In the heatmap, the columns and rows represent the 60 EV samples belonging to the nine tumor types marked with different colors and the 172 proteins, respectively. Both the columns and rows are clustered. Dendrogram branches ending in a square indicate the elements to be included in the Train set. **b** t-SNE plot of the selected 172 proteins. The dots with different colors represent the 60 individual EV samples belonging to the nine tumor types. In the plot, the color gradient indicates the dot density. **c** Confusion matrix of the classification results using the selected proteins on the Test set. Each row of the matrices represents the instances in an actual class while each column represents the instances in a predicted class. Diagonally, the percentage of the correct classification is shown in blue. The percentage of errors is indicated in red

proliferation capacity in the Train set, respectively (invasion panel and proliferation panel).

The Test set was then used to validate the predictive value of the panels using multiple linear regression.

Multiple linear regression showing significant results for both the invasion panel and the proliferation panel ($p<0.0001$), we also obtained remarkably high coefficients of determination: $R^2=0.68$ for the invasion, $R^2=0.62$ for the proliferation capacity (Fig. 4). Pooling the Test and Train sets, the R^2 values were found to be 0.71 and 0.69, respectively.

After validation on the Test set confirmed the predictive value of the proteins, both of the 20- and 15-member panels (Additional file 4) were then subjected to hierarchical clustering, which resulted in 2–2 clusters (Fig. 5): one cluster that appears to be negatively correlated and another that appears to be positively correlated with invasion or proliferation capacity.

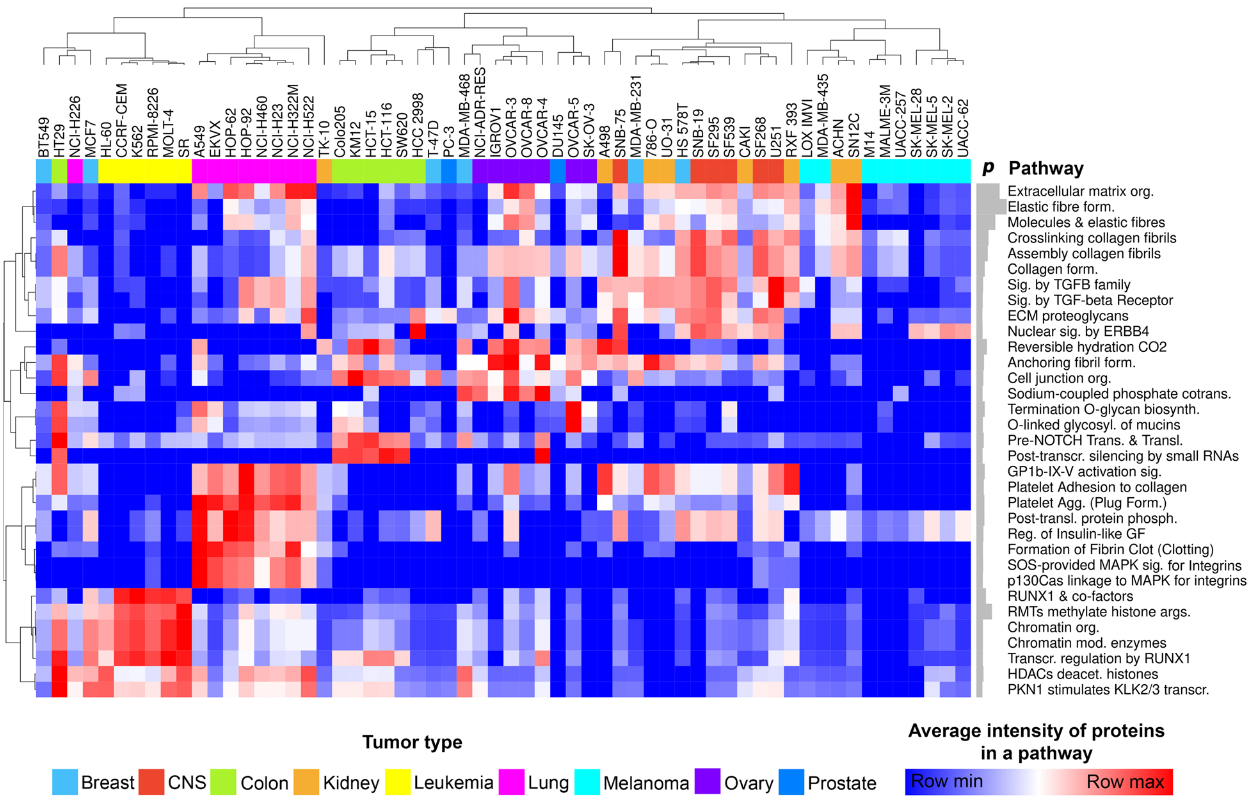


Fig. 3 Biological signaling pathways affected by the 172 selected proteins of the discriminative protein panel. The columns marked with different colors represent the 60 EV samples, while the rows indicate the various signaling pathways. Both the 60 samples and pathways were clustered hierarchically. The heatmap values represent the average intensity of the proteins that are part of a given signal pathway. The gray barplots next to the names of the pathways indicate the -log$_{10}$(p value). In all instances, $p < 0.05$. (agg.: aggregation; biosynth.: biosynthesis; cotrans.: cotransporters; deacet.: deacetylate; form.: formation; mod.: modifying; org.: organization; phosph.: phosphorylation; prots.: proteoglycans; sig.: signaling; trans.: transcription; transl.: translocation)

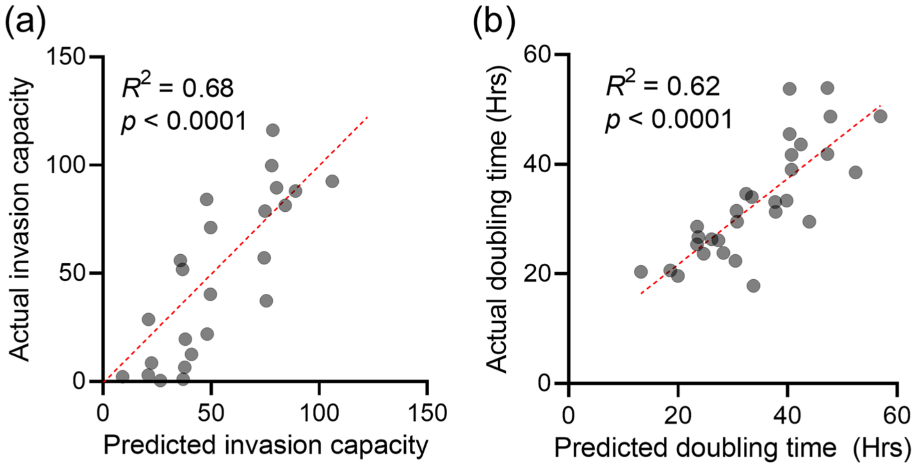


Fig. 4 Results of the multiple linear regression. **a** Multiple linear regression of invasion capacity. The invasion capacity predicted by the invasion panel for each sample in the Test set is plotted on the x-axis, while the actual invasion capacity is plotted on the y-axis. **b** Multiple linear regression of proliferation capacity. The doubling time predicted by the invasion panel for each sample in the Test set is plotted on the x-axis, while the actual doubling time is plotted on the y-axis. (R^2—coefficient of determination; p—p value.)

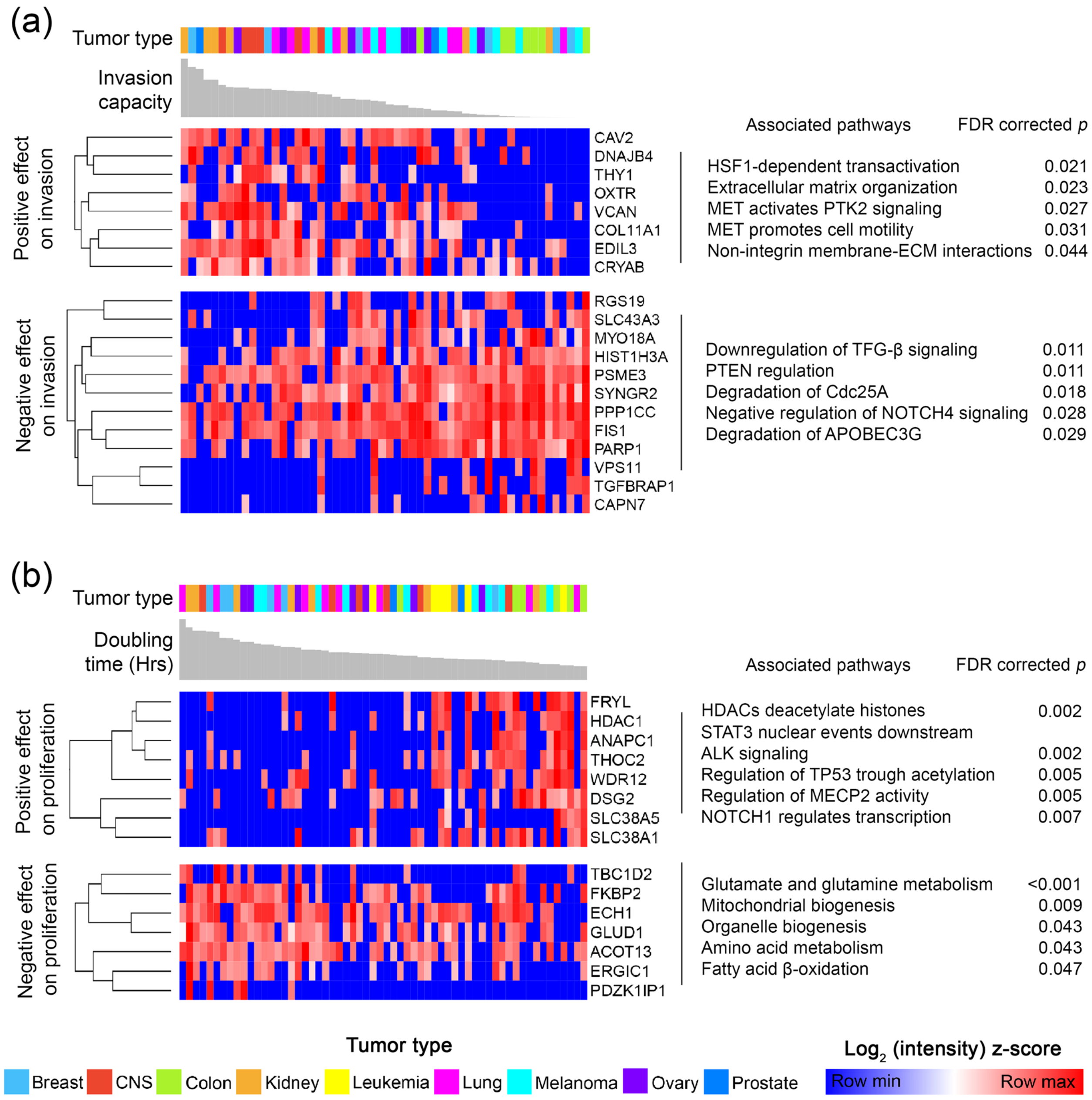


Fig. 5 Predictive proteins for invasion and proliferation capacity. **a** Predictive protein panel for invasion capacity (invasion panel). The columns marked with different colors and the gray barplots indicate the 54 EV samples with the invasion capacity measured for the cell line of origin (leukemia not included). The rows indicate the proteins, which were clustered hierarchically. Two defined clusters were separated from each other. **b** Predictive protein panel for proliferation capacity (proliferation panel). The columns marked with different colors and the gray barplots indicate the 60 EV samples with the doubling time (in hours) measured for the cell line of origin. The rows indicate the proteins, which were clustered hierarchically. Two defined clusters were separated from each other. It should be noted that higher doubling time means lower proliferation capacity as it indicates more time for cell division

Of the 20-member invasion panel, eight proteins (CAV2, DNAJB4, THY1, OXTR, VCAN, COL11A1, EDIL3, CRYAB) positively predicted the invasion capacity of the cell lines. Based on Reactome pathway analysis, these proteins were significantly associated with signaling pathways that upregulate tumor cell maintenance, invasion and binding to the extracellular matrix. Similarly, the enrichment analysis of the remaining twelve proteins that negatively predict invasion capacity was consistent with the regression results:

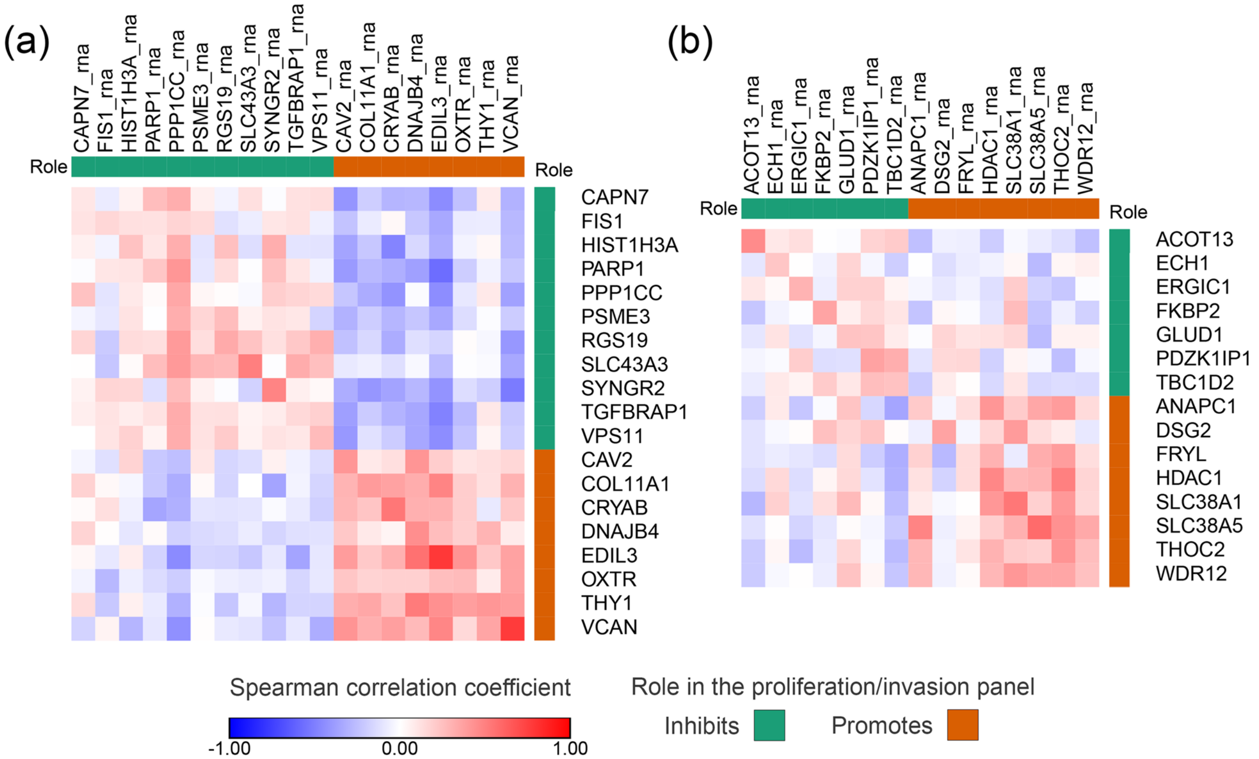


Fig. 6 Correlation of EV protein and cellular RNA content. The heatmaps show the correlation between cellular RNAs and EV proteins of invasion (**a**) and proliferation (**b**) panel members. Columns represent the cellular RNA, rows represent the EV proteins

these proteins play a role in pathways that negatively regulate the invasion (Fig. 5a).

The eight proteins that positively influence proliferation capacity were associated with processes linked to cell cycle. While seven proteins negatively associated with proliferation are linked to metabolic pathways (Fig. 5b).

We further attempted to gain more support for our invasion and proliferation capacity prediction panels by examining their impact on patients' survival time.

The Human Protein Atlas (HPA) was considered an appropriate database for this purpose, as it contains survival times for a large number of cancer patients for all nine cancer types and is easily accessible. However, we had to take into account the limitation that HPA contains tissue RNA expression data and not EV proteomic data.

Accordingly, before utilizing the HPA database, we had to assess the similarity of EV protein and cellular RNA patterns to be permitted to investigate the effect of in vivo RNA tissue expression of panel members on survival time.

First, we examined how the EV protein panels (invasion and proliferation) and the cellular RNAs correlate with each other (Fig. 6). Based on the results, the RNA and protein patterns of the invasion panel showed a moderately strong concordance ($R_V = 0.51$, $p = 0.020$). While a weaker but still significant relationship was observed when comparing the RNA and protein matrices of the proliferation panel ($R_V = 0.39$, $p = 0.048$). Notably, we observed stronger pairwise correlations between protein and RNA content for the promoting members of both panels.

After assessing the relationship between EV protein and cellular RNA pattern, we attempted to use the cellular RNA to estimate the invasion and proliferation capacity of cells using the panel members.

Based on the cellular RNA, invasion capacity could be estimated at $R^2 = 0.77$ ($p < 0.0001$) and proliferation capacity at $R^2 = 0.32$ ($p = 0.037$).

The in vitro data suggested that the EV proteomic and cellular RNA patterns are in concordance and that the cellular RNA content is also related to invasion and proliferation capacity in a similar way as the EV proteome. This prompted us to investigate the impact of in vivo RNA tissue expression of panel members on patient survival.

Using the HPA database, we collected clinical data on the tissue expression of our panel members in the nine tumor types from 4,665 patients, then examined the

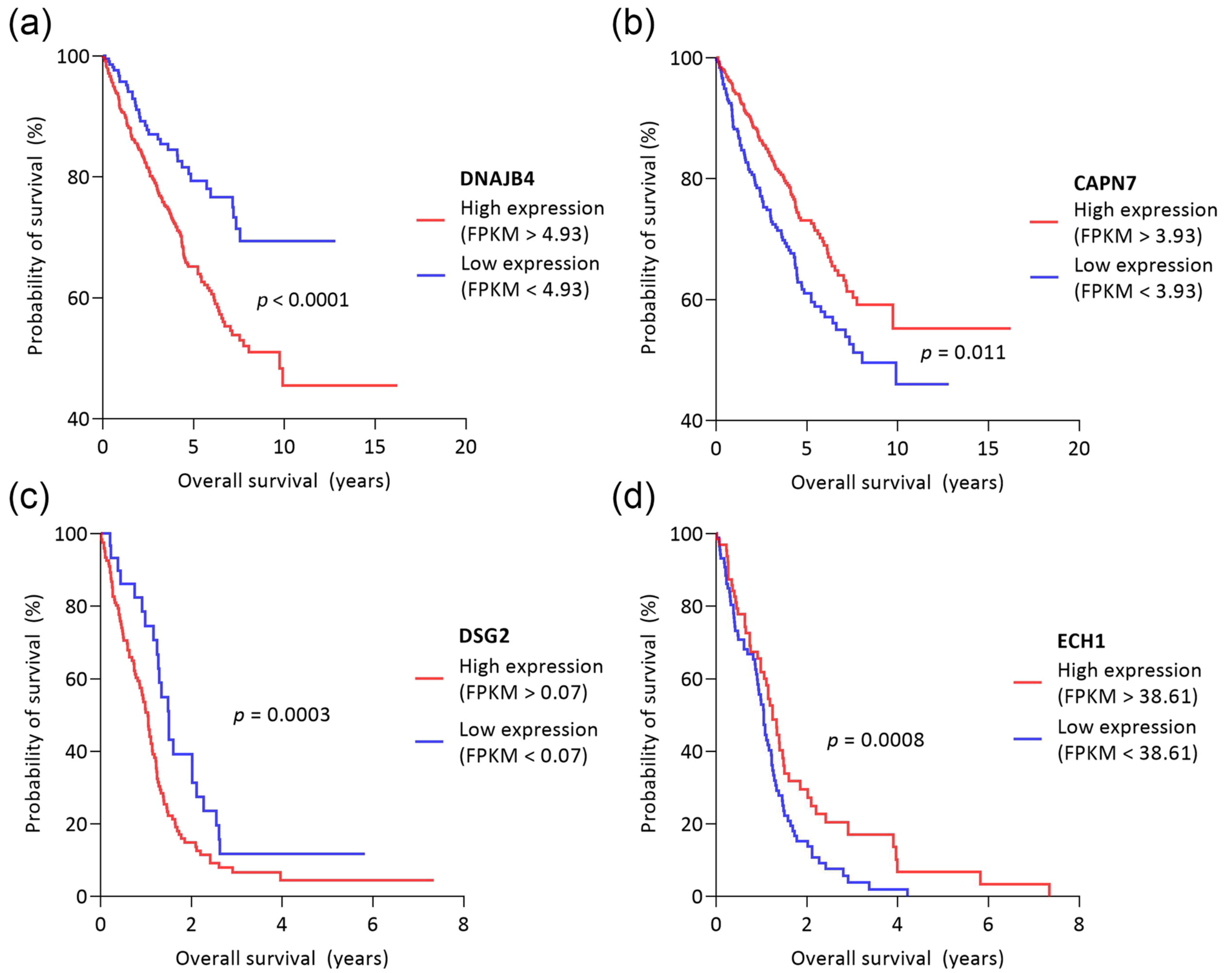


Fig. 7 Survival functions for different expression levels of DNAJB4, CAPN7, DSG2, ECH1. The figure shows 4 exemplary proteins selected from the members of the invasion and proliferation panel and their impact on patients' survival. **a** DNAJB4, which we found to be positively associated with invasion and which the Human Protein Atlas (HPA) suggests that its high expression is associated with a worse prognosis in kidney tumors ($n = 877$). **b** CAPN7 protein, which in our study is negatively associated with invasion and which the HPA suggests may be associated with a favorable prognosis in kidney tumors. **c** DSG2 protein which in our study positively predicted the proliferation capacity is a negative prognostic factor in CNS tumors, based on HPA. **d** Based on our results, ECH1 protein negatively predicted the proliferation capacity, and it is a favorable prognostic marker for CNS tumors

relationship between tissue expression and 5-year survival rate.

In the HPA database, tissue expression was found for 19 of the 20 proteins of the invasion panel (Additional file 5).

According to the HPA, high expression of CAV2, COL11A1, DNAJB4, THY1 and VCAN decreased the 5-year survival for breast, CNS, colon, kidney, lung and ovarian tumors (Fig. 7a). These findings are in line with our results, as these proteins were found to be positively associated with invasion capacity according to multiple linear regression analysis.

The CRYAB protein was found to be controversial, as our results showed a positive association with invasion, but in HPA, high tissue expression was associated with a better prognosis in CNS tumors. Nevertheless, in colon tumors, high expression was a negative prognostic marker.

The case is similar for EDIL3, which is positively associated with invasion capacity according to multiple linear regression analysis, but based on the HPA, higher tissue expression is associated with better 5-year survival in colon tumors. However, it still was a significantly worse prognostic marker in breast, kidney and melanoma patients.

Overall, the effects on survival found in the HPA database and the effect of the proteins on invasion capacity as determined in our study were consistent in 90% of the cases.

Based on multiple linear regression, twelve proteins in our study were found to be negatively correlated with invasion capacity. Comparing this finding to the HPA database, we found more inconsistencies: according to the HPA, the twelve proteins are favored prognostic markers for 5-year survival in most cases (73.18%) (Fig. 7b), but in 26.82%, the proteins have an adverse effect on survival than the expected. For example, HIST1H3A showed a negative association with invasiveness in our study, but its high expression negatively affected the survival rate of CNS tumor patients according to the HPA database (Additional file 5).

Tissue expression was found for all the 15 proteins of the proliferation panel (Additional file 6). The proliferation panel contains seven proteins which were found to negatively predict the proliferation capacity. According to HPA, high tissue expression of these seven proteins significantly increased the 5-year survival in 64.71% of cases (Fig. 7c). Vice versa, the high expression of the eight proteins which positively predict the proliferation capacity significantly reduces the 5-year survival in 72.41% of cases (Fig. 7d).

Taken as a whole, the EV proteome and in vitro cellular RNA pattern of the panel members showed concordance, and the effect of in vivo tissue RNA expression of the panel members on patient survival is consistent with the results of our linear regression model. The finding potentially suggests the involvement of invasion and proliferation panels in the tumorous processes.

It is noteworthy that the inconsistency with HPA appears for those variables where the in vitro EV proteome and cellular RNA pattern did not show a strong correlation (invasion capacity inhibitory members) (Fig. 4), or cellular RNA did not prove to be a sufficient predictor (overall the proliferation panel).

Discussion

Nowadays, EVs are considered as a novel and promising tool for liquid biopsy-based cancer diagnosis, prognosis and therapeutic decisions. However, there are barely explored segments of their potential clinical applicability.

In the present study, we aimed to determine the degree of specificity of the proteome carried by EVs from various tumor types, as well as whether the EVs' molecular pattern can be used to predict the invasion capacity and proliferation rate of the donor cells.

In our meta-analyses, we investigated the proteome of EVs isolated from the supernatant of NCI-60 cell lines. Of the total proteome, 213 proteins were present in all EV samples (core proteome). Although these proteins were observed in all tumors, they showed some degree of specificity.

Based on Gene Ontology Enrichment Analysis, these protein sets are associated with biological pathways, molecular functions, and cellular components including protein targeting, cotranslational modifications, RNA binding and processing, ribosomal subunit, and exocytotic pathways. These findings are consistent with those previously described by Hurwitz et al. [41, 54]. As it has been pointed out before, this enrichment may indicate that the core proteome facilitates cell-to-cell communication by directly translating the mRNA content of EVs following fusion with the target cell.

Even though the core proteome showed differences between the nine tumor types, the reason for these differences could not be determined from the available data. Our correlation analyses suggested that the distinct core proteome pattern was not caused by the difference of EV production rates or EV type between the nine tumor types. Therefore, we assumed that the source of the observed variance in the core proteome is the different origin of the nine tumor types.

Extending the analysis to the entire proteome, then to the selected protein set significantly improved classification accuracy, indicating that the molecular signature carried by EVs is remarkably characteristic of certain tumor types, and this specificity could be further increased by using the appropriate selection methods.

This finding is in accordance with previous literature data. However, most studies have attempted to distinguish between cancerous samples and matched controls, or to subcategorize different tumor types in both in vivo and in vitro experiments [16].

For example, by selecting the proteins detected in EVs, Vinik et al. showed that the control and breast cancer patient groups were significantly distinguishable from each other [17].

The diagnostic efficacy of vesicles has also been demonstrated for brain tumors. In an in vivo experiment with mice, Anastasi et al. used principal component analysis to show that the proteome of control and mice with glioblastoma multiforme differed significantly [18].

Moreover, diagnostic importance has also been reported for ovarian, colon cancer and leukemia [19–21].

In addition to distinguishing a tumor cohort from a matched control sample, studies can be found about stratifying a cancerous disease according to different characteristics. For example, Li et al. investigated plasma EVs to highlight leukemia patient groups with different imatinib resistance [22]. Choi et al. distinguished between primary and metastatic colon tumors [23]. Mallawaaratchy et al. identified glioblastoma subtypes of aggressiveness [24], and Rontogianni et al. pointed out that proteomic analysis of EVs allows the differentiation of breast cancer subtypes [25].

Our study differs from these in that our aim was not to investigate the differences from control samples or to subcategorize a certain tumor, but to distinguish a wide range of tumors with different tissue origin. In a well-written article, which was the source of the NCI-60 proteomic data set Hurwitz et al. have already demonstrated that some tumor types are distinguishable from the others [41].

Approaching this valuable dataset with the evolving machine learning based classifier algorithms suggests that the proteomic content carried by cancer EVs is more specific than expected and previously reported.

Uncovering tumor-specific signaling pathways is a key element in identifying drug targets [55]. Most research focuses on the analysis of tissue, however, obtaining tissue biopsy from certain tumors, particularly brain tumors, carries high risks for the patient, has limited reproducibility, and does not provide reliable information due to intratumoral heterogeneity [56]. However, these challenges can be overcome by using EVs isolated from the circulation, as their molecular content provides information about the entire tumorous condition [57].

Although there is a growing body of research on the use of EVs as drug carriers, no studies have investigated the molecular content of EVs in an attempt to identify drug targets [58].

Our results suggest that the proteins showing the largest group differences between the nine tumor types may indicate tumor type-specific signaling pathways and specific strategies.

For example, matrix-related processes were proven to be specifically involved in CNS and kidney tumors. Pointer et al. have shown that collagen matrix structure plays a significant role in the survival of patients with glioblastoma: the presence of disorganized fibers is associated with a significantly worse prognosis [59]. Similar results have been described in kidney cancer, where collagen matrix structure predicted the tumor grade [60].

NOTCH signaling was found to be specifically characteristic for colon cancers based on the EV proteome. Consistent with our findings, several studies have highlighted that NOTCH signaling is essential for the initiation of colon cancer cell development [61].

We also found a strong association between the leukemia EV proteome and processes associated with the transcription factor RUNX1, whose mutation has been shown to play an important role in the development of hematological malignancies [62].

In addition to the above examples, the results of our enrichment study are supported by further literature on leukemia [63], melanoma [64], lung [65, 66] and ovarian cancer [67, 68].

Extending and applying our knowledge on the invasiveness and proliferation rate of cancer cells is vital for the proper treatment and prognosis of patients. In estimating patient survival, the number of metastatic nodules and the size of the tumor mass are particularly crucial variables [69–72].

Our findings suggest that the EV proteome can provide information about the donor cells' proliferation rate, and invasion capacity, which are crucial steps in tumor progression and metastasis formation [73].

The predictive invasion and proliferation panel were subjected to Reactome pathway analysis to reveal the physiological mechanisms of the predicted effects. For instance, we found that EV proteins detected in high invasion capacity tumor cell lines may induce HSF1-dependent transactivation. This finding is supported by literature data; amplification of HSF1 was shown in a wide variety of tumors with a 10.33–26.54% alteration frequency in the most aggressive tumors, i.e. ovarian epithelial tumors, breast cancer, pancreatic cancer [74, 75].

As HSF-1 is a main transactivator of HSPs expression, including HSP60, HSP70, and HSP90, it has multiple effects on cancer progression, such as promoting invasion and metastasis [76].

Our data show that proteins predicting low invasion may cause downregulation of TGF-β signaling. Indeed, TGF-β may function as a tumor promoter by stimulating epithelial-mesenchymal transition (EMT) of tumor cells leading to metastasis [77]. Also, inactivation of TGF-β signaling suppress prostate cancer bone metastasis [78].

Panel members, which positively predict proliferation capacity are significantly associated with reversible histone acetylation by HDAC enzymes. Several studies have investigated HDAC and proliferation; for example, HDAC enzymes are important in melanoma tumor cell proliferation [79]. And again, inhibition of HDACs represses proliferation of head and neck squamous cell carcinoma cells [80]. In addition, various phases of preclinical trials are addressing the inhibition of HDAC in subjects with mutated advanced and unrespectable melanoma (ClinicalTrials.gov ID: NCT02836548, NCT02032810).

From the list of proteins which are associated with lower proliferation, the GLUD1 (glutamate dehydrogenase 1) were shown to influence glutamate and glutamine metabolism. It is evidenced so far that glutamine metabolism enhances the proliferation and tumor growth [76]. However, high expression of GLUD1 may predict good overall patient outcome [81]. Coloff et al. showed negative correlation between GLUD1 and proliferation, concluding that highly proliferative tumors couple glutamine anaplerosis to non-essential amino acid synthesis [82].

Despite the fact that the results of the meta-analysis appear to be supported by other findings, it is important to draw attention to the limitations of our work.

The data set is relatively small compared to the number of elements required for machine learning: it contains proteomic data from EV samples of 60 cell lines, and the nine tumor types have different sample numbers.

However, we found that even 50% of the data was enough for the Train set to learn important patterns from the data that could be applied to the Test set. We believe that despite the small number of elements, we could find generalizable differences. Nevertheless, we acknowledge the importance of validating the findings on a larger dataset to ensure the robustness of the results.

Hurwitz et al. described a strong correlation between the proteomic pattern of EVs and the cellular RNA content [41]. Our study has highlighted that within the entire proteome, our invasion and proliferation panels are also in concordance with the cellular RNA pattern. This finding prompted us to investigate the impact of in vivo RNA expression of panel members on tumor patient survival.

The predictive value of the invasion and proliferation panel established in this study was supported by the literature and the Human Proteome Atlas (HPA) database. Nevertheless, the authors acknowledge and strongly emphasize that comparing the in vitro EV proteome and in vivo tissue RNA expression is an implicit approach even if the relationship between the EV proteome and the in vitro cellular RNA pattern has been successfully assessed. The comparison is not intended to validate the panel members, but rather to suggest potential biomarker targets that may be worthy of further research.

The main limitation of the study is that its results are based on 2D in vitro data. 2D cultures have several limitations, such as perturbation of interactions between the cellular and extracellular environment, changes in cell morphology, polarity and proliferation mode [83]. The authors certainly acknowledge the need for further validation, and consider the results presented here only as promising research candidates, not as an unimprovable approach to the in vivo phenomenon.

A previous meta-analysis has already analyzed the proteome of NCI-60 EVs, but with different assumptions [84]. In this research, the investigation aimed to determine the potential support of EV proteomes in facilitating the functional transfer of cancer hallmarks. The study conducted a meta-analysis, where a comparison was made between EVs and entire cell proteomes derived from the NCI-60 cell lines. A distinct subset of proteins within each cancer hallmark signature was identified, demonstrating both high abundance and consistent expression within EVs across all cell lines.

To our knowledge, ours is the first study to classify such a large number of tumor types based on proteomic data from EVs, looking for discriminative patterns, and to investigate the predictive value for donor cell invasion capacity and proliferation rate using machine learning techniques, which could greatly help in evaluating the potential clinical applications of EVs.

Conclusions

Our results suggest that the extensive body of knowledge on EV omics research to date is worth re-exploring with the emerging and increasingly available state-of-the-art methods. Integrating proteomic data from EVs from different tumor types with cell physiological and clinical data can help to reveal the full potential of EVs in oncology. By studying their molecular content, it may be possible to obtain information on tumor properties that are crucial for patient treatment, such as invasion and proliferation capacity. In addition, they may also allow us to unravel the signaling pathways and biological processes underlying the specific characteristics of different tumor types, helping to identify potential drug targets.

Abbreviations

ANOVA	Analysis of variance
AUC	Area Under the Curve
CAPN7	Calpain 7
CAV2	Caveolin 2
CNS	Central nervous system
COL11A1	Collagen type XI alpha 1 chain
CRYAB	Alpha-crystallin B
DNAJB4	DnaJ heat shock protein family (Hsp40) member B4
DSG2	Desmoglein 2
ECH1	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase
EDIL3	EGF like repeats and discoidin domains 3
ERB4	Erb-b2 receptor tyrosine kinase 4
EV	Extracellular vesicle
FDR	False discovery rate
GLUD1	Glutamate dehydrogenase 1,
HDAC	Histone deacetylases
HIST1H3A	Histone H3.1
HPA	Human Proteome Atlas
HSF1	Heat shock transcription factor 1
HSP	Heat shock protein
LASSO	Least Absolute Shrinkage and Selection Operator
mRNA	Messenger ribonucleic acid
NCI-60	National Cancer Institute 60
NOTCH	Neurogenic locus notch homolog
OXTR	Oxytocin receptor
p	p value
PCA	Principal component analysis
R^2	Coefficient of determination
RNA	Ribonucleic acid
RUNX1	RUNX family transcription factor 1
TGF-β	Transforming growth factor beta
THY1	Thy-1 cell surface antigen
t-SNE	T-distributed stochastic neighbor embedding
VCAN	Versican

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12964-023-01344-5.

Additional file 1. Gene Ontology Enrichment for the entire proteome.

Additional file 2. Gene Ontology Enrichment for the core proteome.

Additional file 3. The selected 172 proteins.

Additional file 4. Members of the invasion and proliferation panels.

Additional file 5. Comparison of the invasion panel with the Human Protein Atlas database.

Additional file 6. Comparison of the proliferation capacity panel with the Human Protein Atlas database.

Acknowledgements
The authors thank Lilla Pintér for her tireless technical assistance.

Authors' contributions
M.B: Data curation, formal analysis, investigation, methodology, visualization, writing – original draft preparation. G.D: writing- original draft preparation, review and editing. E.Gy-S: Verification, writing- review and editing. T.B: writing- review and editing. M.M.K: Writing- review and editing. D.G.M: Resources. P.H: Resources. K.B: Funding acquisition, resources, writing – review and editing. M.H: Conceptualization, investigation, methodology, project administration, supervision, writing – review and editing.

Funding
Open access funding provided by ELKH Biological Research Center. This study was supported by the following research grants: TKP2021-EGA-09 (P.H., K.B.); the Albert Szent-Györgyi Research Fund provided by the University of Szeged (K.B.), OTKA-K143255 (K.B.); ÚNKP22-3 – New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund (M.B.).

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Immunology, Albert Szent-Györgyi Medical School, Faculty of Science and Informatics, University of Szeged, 6726 Szeged, Hungary. [2]Doctoral School of Interdisciplinary Medicine, Albert Szent-Györgyi Medical School, University of Szeged, 6720 Szeged, Hungary. [3]Laboratory of Microscopic Image Analysis and Machine Learning, Institute of Biochemistry, Biological Research Centre, Hungarian Research Network (HUN-REN), Szeged 6726, Hungary. [4]Department of Biomedical Sciences, Florida State University College of Medicine, Tallahassee, FL 32306, USA.

Received: 17 August 2023 Accepted: 27 September 2023
Published online: 20 November 2023

References
1. Chakraborty S, Hosen MdI, Ahmed M, Shekhar HU. Onco-Multi-OMICS approach: a new frontier in cancer research. BioMed Res Int. 2018;2018:1–14.
2. Heo YJ, Hwa C, Lee GH, Park JM, An JY. Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. Mol Cells. 2021;44(7):433–43.
3. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021;19:949–60.
4. Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. J Natl Cancer Inst. 2013;105(7):452–8.
5. Sarhadi VK, Armengol G. Molecular Biomarkers in Cancer. Biomolecules. 2022;12(8):1021.
6. Martins I, Ribeiro IP, Jorge J, Gonçalves AC, Sarmento-Ribeiro AB, Melo JB, et al. Liquid biopsies: applications for cancer diagnosis and monitoring. Genes. 2021;12(3):349.
7. Ciferri MC, Quarto R, Tasso R. Extracellular vesicles as biomarkers and therapeutic tools: from pre-clinical to clinical applications. Biology. 2021;10(5):359.
8. Yáñez-Mó M, Siljander PRM, Andreu Z, BedinaZavec A, Borràs FE, Buzas EI, et al. Biological properties of extracellular vesicles and their physiological functions. J Extracell Vesicles. 2015;4(1):27066.
9. Théry C, Witwer KW, Aikawa E, Alcaraz MJ, Anderson JD, Andriantsitohaina R, et al. Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. J Extracell Vesicles. 2018;7(1):1535750.
10. Vardaki I, Ceder S, Rutishauser D, Baltatzis G, Foukakis T, Panaretakis T. Periostin is identified as a putative metastatic marker in breast cancer-derived exosomes. Oncotarget. 2016;7(46):74966–78.
11. Melo SA, Luecke LB, Kahlert C, Fernandez AF, Gammon ST, Kaye J, et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. Nature. 2015;523(7559):177–82.
12. Boukouris S, Mathivanan S. Exosomes in bodily fluids are a highly stable resource of disease biomarkers. Prot Clin Appl. 2015;9(3–4):358–67.
13. Dobra G, Bukva M, Szabo Z, Bruszel B, Harmati M, Gyukity-Sebestyen E, et al. Small extracellular vesicles isolated from serum may serve as signal-enhancers for the monitoring of CNS tumors. IJMS. 2020;21(15):5359.
14. Bukva M, Dobra G, Gomez-Perez J, Koos K, Harmati M, Gyukity-Sebestyen E, et al. Raman spectral signatures of serum-derived extracellular vesicle-enriched isolates may support the diagnosis of CNS tumors. Cancers. 2021;13(6):1407.
15. Zhou B, Xu K, Zheng X, Chen T, Wang J, Song Y, et al. Application of exosomes as liquid biopsy in clinical diagnosis. Sig Transduct Target Ther. 2020;5(1):144.
16 Liu SYA, Liao Y, Hosseinifard H, Imani S, Wen QL. Diagnostic role of extracellular vesicles in cancer: a comprehensive systematic review and meta-analysis. Front Cell Dev Biol. 2021;15(9):705791.
17. Vinik Y, Ortega FG, Mills GB, Lu Y, Jurkowicz M, Halperin S, et al. Proteomic analysis of circulating extracellular vesicles identifies potential markers of breast cancer progression, recurrence, and response. Sci Adv. 2020;6(40):eaba5714.
18. Anastasi F, Greco F, Dilillo M, Vannini E, Cappello V, Baroncelli L, et al. Proteomics analysis of serum small extracellular vesicles for the longitudinal study of a glioblastoma multiforme mouse model. Sci Rep. 2020;10(1):20498.
19. Lai H, Guo Y, Tian L, Wu L, Li X, Yang Z, et al. Protein panel of serum-derived small extracellular vesicles for the screening and diagnosis of epithelial ovarian cancer. Cancers. 2022;14(15):3719.
20. Lee CH, Im EJ, Moon PG, Baek MC. Discovery of a diagnostic biomarker for colon cancer through proteomic profiling of small extracellular vesicles. BMC Cancer. 2018;18(1):1058.
21. Zhu S, Xing C, Li R, Cheng Z, Deng M, Luo Y, et al. Proteomic profiling of plasma exosomes from patients with B-cell acute lymphoblastic leukemia. Sci Rep. 2022;12(1):11975.
22. Li MY, Zhao C, Chen L, Yao FY, Zhong FM, Chen Y, et al. Quantitative proteomic analysis of plasma exosomes to identify the candidate biomarker of imatinib resistance in chronic myeloid leukemia patients. Front Oncol. 2021;21(11):779567.

23. Choi DS, Choi DY, Hong B, Jang S, Kim DK, Lee J, et al. Quantitative proteomics of extracellular vesicles derived from human primary and metastatic colorectal cancer cells. J Extracell Vesicles. 2012;1(1):18704.
24. Mallawaaratchy DM, Hallal S, Russell B, Ly L, Ebrahimkhani S, Wei H, et al. Comprehensive proteome profiling of glioblastoma-derived extracellular vesicles identifies markers for more aggressive disease. J Neurooncol. 2017;131(2):233–44.
25. Rontogianni S, Synadaki E, Li B, Liefaard MC, Lips EH, Wesseling J, et al. Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. Commun Biol. 2019;2(1):325.
26. Weinstein JN. Integromic analysis of the NCI-60 cancer cell lines. Green JE, editor. Breast Dis. 2004;19(1):11–22.
27. Weinstein JN. Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Mol Cancer Ther. 2006;5(11):2601–5.
28. Gholami AM, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, et al. Global proteome analysis of the NCI-60 cell line panel. Cell Rep. 2013;4(3):609–20.
29. Sinha A, Ignatchenko V, Ignatchenko A, Mejia-Guerrero S, Kislinger T. In-depth proteomic analyses of ovarian cancer cell line exosomes reveals differential enrichment of functional categories compared to the NCI 60 proteome. Biochem Biophys Res Commun. 2014;445(4):694–701.
30. Staubach S, Razawi H, Hanisch FG. Proteomics of MUC1-containing lipid rafts from plasma membranes and exosomes of human breast carcinoma cells MCF-7. Proteomics. 2009;9(10):2820–35.
31. Ji H, Greening DW, Barnes TW, Lim JW, Tauro BJ, Rai A, et al. Proteome profiling of exosomes derived from human primary and metastatic colorectal cancer cells reveal differential expression of key metastatic factors and signal transduction components. Proteomics. 2013;13(10–11):1672–86.
32. Keerthikumar S, Gangoda L, Liem M, Fonseka P, Atukorala I, Ozcitti C, et al. Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. Oncotarget. 2015;6(17):15375–96.
33. Valenzuela MMA, Ferguson Bennit HR, Gonda A, Diaz Osterman CJ, Hibma A, Khan S, et al. Exosomes secreted from human cancer cell lines contain Inhibitors of Apoptosis (IAP). Cancer Microenvironment. 2015;8(2):65–73.
34. Kong JN, He Q, Wang G, Dasgupta S, Dinkins MB, Zhu G, et al. Guggulsterone and bexarotene induce secretion of exosome-associated breast cancer resistance protein and reduce doxorubicin resistance in MDA-MB-231 cells: ceramide reduces multidrug resistance in breast cancer. Int J Cancer. 2015;137(7):1610–20.
35. Shedden K, Xie XT, Chandaroy P, Chang YT, Rosania GR. Expulsion of small molecules in vesicles shed by cancer cells: association with gene expression and chemosensitivity profiles. Cancer Res. 2003;63(15):4331–7.
36. Clayton A, Mitchell JP, Court J, Linnane S, Mason MD, Tabi Z. Human tumor-derived exosomes down-modulate NKG2D expression. J Immunol. 2008;180(11):7249–58.
37. Sung BH, Ketova T, Hoshino D, Zijlstra A, Weaver AM. Directional cell movement through tissues is controlled by exosome secretion. Nat Commun. 2015;6(1):7164.
38. Webber JP, Spary LK, Sanders AJ, Chowdhury R, Jiang WG, Steadman R, et al. Differentiation of tumour-promoting stromal myofibroblasts by cancer exosomes. Oncogene. 2015;34(3):290–302.
39. Phuyal S, Hessvik NP, Skotland T, Sandvig K, Llorente A. Regulation of exosome release by glycosphingolipids and flotillins. FEBS J. 2014;281(9):2214–27.
40. Kosaka N, Iguchi H, Hagiwara K, Yoshioka Y, Takeshita F, Ochiya T. Neutral sphingomyelinase 2 (nSMase2)-dependent exosomal transfer of angiogenic MicroRNAs regulate cancer cell metastasis. J Biol Chem. 2013;288(15):10849–59.
41. Hurwitz SN, Rider MA, Bundy JL, Liu X, Singh RK, Meckes DG. Proteomic profiling of NCI-60 extracellular vesicles uncovers common protein cargo and cancer type-specific biomarkers. Oncotarget. 2016;7(52):86999–7015.
42. Hurwitz SN, Meckes DG. Extracellular vesicle integrins distinguish unique cancers. Proteomes. 2019;7(2):14.
43. Arjmand B, Hamidpour SK, Tayanloo-Beik A, Goodarzi P, Aghayan HR, Adibi H, et al. Machine learning: a new prospect in multi-omics data analysis of cancer. Front Genet. 2022;27(13):824451.
44. DeLosh RM, Shoemaker RH. Evaluation of Real-Time In Vitro Invasive Phenotypes. In: Stein US, editor. Metastasis. New York, NY: Springer US; 2021. p. 165–80. Methods in Molecular Biology; vol. 2294. https://doi.org/10.1007/978-1-0716-1350-4_12. Cited 2023 Jan 10.
45. Cell Lines in the In Vitro Screen. Available online: https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm. Accessed on 01.10.2023.
46. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, et al. Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Mol Cancer Ther. 2009;8(7):1878–84.
47. The Human Protein Atlas. Available online: https://www.proteinatlas.org/. Accessed on 01.10.2023.
48. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, et al. Orange: data mining toolbox in python. J Mach Learn Res. 2013;14(1):2349–53.
49. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2019;36(8):2628–9.
50. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687–92.
51. Morpheus. Available online: https://software.broadinstitute.org/morpheus. Accessed on 01.10.2023.
52. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the RV- coefficient. Appl Stat. 1976;25(3):257.
53. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;15(1):162.
54. Zhu Y, Chen X, Pan Q, Wang Y, Su S, Jiang C, et al. A comprehensive proteomics analysis reveals a secretory path- and status-dependent signature of exosomes released from tumor-associated macrophages. J Proteome Res. 2015;14(10):4319–31.
55. Paananen J, Fortino V. An omics perspective on drug target discovery platforms. Brief Bioinform. 2020;21(6):1937–53.
56. Russano M, Napolitano A, Ribelli G, Iuliani M, Simonetti S, Citarella F, et al. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. J Exp Clin Cancer Res. 2020;39(1):95.
57. Corvigno S, Johnson AM, Wong KK, Cho MS, Afshar-Kharghan V, Menter DG, et al. Novel markers for liquid biopsies in cancer management: circulating platelets and extracellular vesicles. Mol Cancer Ther. 2022;21(7):1067–75.
58. Hernandez-Oller L, Seras-Franzoso J, Andrade F, Rafael D, Abasolo I, Gener P, et al. Extracellular vesicles as drug delivery systems in cancer. Pharmaceutics. 2020;12(12):1146.
59. Pointer KB, Clark PA, Schroeder AB, Salamat MS, Eliceiri KW, Kuo JS. Association of collagen architecture with glioblastoma patient survival. JNS. 2016;126(6):1812–21.
60. Best SL, Liu Y, Keikhosravi A, Drifka CR, Woo KM, Mehta GS, et al. Collagen organization of renal cell carcinoma differs between low and high grade tumors. BMC Cancer. 2019;19(1):490.
61. Sikandar SS, Pate KT, Anderson S, Dizon D, Edwards RA, Waterman ML, et al. *NOTCH* signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. Can Res. 2010;70(4):1469–78.
62. Sood R, Kamikubo Y, Liu P. Role of RUNX1 in hematological malignancies. Blood. 2017;129(15):2070–82.
63. Zhang J, Gao X, Yu L. Roles of histone deacetylases in acute myeloid leukemia with fusion proteins. Front Oncol. 2021;1(11):741746.
64. Lau C, Killian KJ, Samuels Y, Rudloff U. ERBB4 Mutation Analysis: Emerging Molecular Target for Melanoma Treatment. In: Thurin M, Marincola FM, editors. Molecular Diagnostics for Melanoma. Totowa, NJ: Humana Press; 2014. p. 461–80. Methods Mol Biol; 1102. https://doi.org/10.1007/978-1-62703-727-3_24. Cited 2023 Jan 27.
65. Xu L, Xu F, Kong H, Zhao M, Ye Y, Zhang Y. Effects of reduced platelet count on the prognosis for patients with non-small cell lung cancer treated with EGFR-TKI: a retrospective study. BMC Cancer. 2020;20(1):1152.
66. Królczyk G, Ząbczyk M, Czyżewicz G, Plens K, Prior S, Butenas S, et al. Altered fibrin clot properties in advanced lung cancer: impact of chemotherapy. J Thorac Dis. 2018;10(12):6863–72.
67. Nurgalieva AK, Popov VE, Skripova VS, Bulatova LF, Savenkova DV, Vlasenkova RA, et al. Sodium-dependent phosphate transporter NaPi2b as a potential predictive marker for targeted therapy of ovarian cancer. Biochem Biophys Rep. 2021;28:101104.
68. Klemba A, Bodnar L, Was H, Brodaczewska KK, Wcislo G, Szczylik CA, et al. Hypoxia-mediated decrease of ovarian cancer cells reaction to treatment: significance for chemo- and immunotherapies. IJMS. 2020;21(24):9492.

69. Wu S, Chen JN, Zhang QW, Tang CT, Zhang XT, Tang MY, et al. A new metastatic lymph node classification-based survival predicting model in patients with small bowel adenocarcinoma: a derivation and validation study. EBioMedicine. 2018;32:134–41.
70. Narod SA. Tumour size predicts long-term survival among women with lymph node-positive breast cancer. Curr Oncol. 2012;19(5):249–53.
71. Liu Y, He M, Zuo WJ, Hao S, Wang ZH, Shao ZM. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. Front Oncol. 2021;9(11):585613.
72. Wang J, Cao Z, Wang C, Zhang H, Fan F, Zhang J, et al. Prognostic impact of tumor size on patients with neuroblastoma in a SEER -based study. Cancer Med. 2022;11(14):2779–89.
73. van Zijl F, Krupitza G, Mikulits W. Initial steps of metastasis: cell invasion and endothelial transmigration. Mutat Res. 2011;728(1–2):23–34.
74. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
75. Cyran AM, Zhitkovich A. Heat shock proteins and HSF1 in cancer. Front Oncol. 2022;2(12):860320.
76. Lauber K, Brix N, Ernst A, Hennel R, Krombach J, Anders H, et al. Targeting the heat shock response in combination with radiotherapy: Sensitizing cancer cells to irradiation-induced cell death and heating up their immunogenicity. Cancer Lett. 2015;368(2):209–29.
77. Hao Y, Baker D, ten Dijke P. TGF-β-mediated epithelial-mesenchymal transition and cancer metastasis. IJMS. 2019;20(11):2767.
78. Dai Y, Wu Z, Lang C, Zhang X, He S, Yang Q, et al. Copy number gain of ZEB1 mediates a double-negative feedback loop with miR-33a-5p that regulates EMT and bone metastasis of prostate cancer dependent on TGF-β signaling. Theranostics. 2019;9(21):6063–79.
79. Reichert N, Choukrallah MA, Matthias P. Multiple roles of class I HDACs in proliferation, differentiation, and development. Cell Mol Life Sci. 2012;69(13):2173–87.
80. Kakiuchi A, Kakuki T, Ohwada K, Kurose M, Kondoh A, Obata K, et al. HDAC inhibitors suppress the proliferation, migration and invasiveness of human head and neck squamous cell carcinoma cells via p63-mediated tight junction molecules and p21-mediated growth arrest. Oncol Rep. 2021;45(4):46.
81. Craze ML, El-Ansari R, Aleskandarany MA, Cheng KW, Alfarsi L, Masisi B, et al. Glutamate dehydrogenase (GLUD1) expression in breast cancer. Breast Cancer Res Treat. 2019;174(1):79–91.
82. Coloff JL, Murphy JP, Braun CR, Harris IS, Shelton LM, Kami K, et al. Differential glutamate metabolism in proliferating and quiescent mammary epithelial cells. Cell Metab. 2016;23(5):867–80.
83. Kapałczyńska M, Kolenda T, Przybyła W, Zajączkowska M, Teresiak A, Filas V, et al. 2D and 3D cell cultures – a comparison of different types of cancer cell cultures. Archives of Medical Science. Termedia Sp. z.o.o.; 2016.https://doi.org/10.5114/aoms.2016.63743.
84. Matthiesen R. Extra-cellular vesicles carry proteome of cancer hallmarks. Front Biosci. 2020;25(3):398–436.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.