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

Thesis booklet

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 related publications:

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)**

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: 60.69

1. INTRODUCTION

The following thesis booklet contains only the essential parts of the thesis. 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

1.2. Extracellular vesicles in cancer

Extracellular vesicles (EVs) are lipid bilayered particles secreted into the extracellular space by all living cells and can be classified into two basic types: exosomes and ectosomes. Exosomes are of endosomal origin and are formed by invagination of endosomal membranes. This process leads to the formation of multivesicular bodies (MVBs) enclosing intraluminal vesicles. The resulting MVBs fuse with the cell membrane, releasing the intraluminal vesicles they carry. The vesicles thus released into the extracellular space are henceforth referred to as exosomes. Ectosomes, on the other hand, are EVs derived directly from the plasma membrane and are formed by outward membrane budding.

However, definitive molecular markers of the different biogenetic 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.

Although EVs may differ in their pathway of biogenesis, their molecular composition is similar: they contain lipids, proteins, nucleic acids and low molecular weight metabolites. In addition to the molecules found in all EVs, they also carry cell-type specific components that reflect the characteristics of the donor cells (cells that release EVs).

Their overall importance lies in their ability to transmit information between donor cells and recipient cells (cells that come into contact with EVs) through the molecular cargo. Communication via EVs regulates both normal physiological processes (e.g. pregnancy, immune response, neuronal development, angiogenesis, blood clotting) and pathological processes (e.g. tumours, neurodegenerative pathologies).

1.3 Extracellular vesicles in biomarker research

To highlight the potential of EVs for a wide range of healthcare applications, including oncotherapy, vaccination, immunomodulatory or regenerative therapies, and drug delivery, many clinical studies have been conducted to date. In particular, EVs could play a key role in liquid biopsy-based future biomarker research. Previous published studies have shown that cancer patients have significantly higher concentrations of EVs in their blood compared to healthy subjects, derived from tumour tissue and tumour-associated immune processes. Therefore, EVs isolated from serum or plasma samples may be a rich source of biomarkers specific for the tumorous condition.

1.4. Approaches in biomarker research

Two approaches are currently used to investigate the molecular content carried by EVs: (I) the analysis of individual molecular types by omics methods (e.g. proteomics, lipidomic, genomics), and (II) the characterization of the overall molecular complexity of EVs by vibrational spectroscopy (e.g. Raman spectroscopy).

The first method is well standardized and has proven its success in many studies when it comes to the identification of specific molecules, but in some cases it has limitations. For example, results from different omics approaches often do not overlap, which can lead to a lack of understanding of cancer biology. Examples of this include tumor diseases that show markedly high variability between patients (e.g. glioblastoma multiforme, GBM). Therefore, a particular protein or nucleic acid biomarker that is clinically meaningful in our patient sample may lose its relevance when testing new samples.

The second approach, the use of vibrational spectroscopy such as Raman spectroscopy, offers a new and promising way to analyze the full molecular composition of EVs. Raman spectroscopy can characterize the total molecular content of a sample based on the phenomenon of light scattering. Simultaneous analysis of the complete molecular composition allows not to limit the analysis to a single type of molecule; thus, a more generalized diagnostic model can be created.

1.5 The increasing prevalence of machine learning in cancer research

Machine learning — a branch of artificial intelligence — is becoming increasingly popular in cancer research because of its ability to analyze complex and large data sets. Traditional statistical methods have struggled to handle high-dimensional and complex data. Machine learning algorithms, on the other hand, "autonomously" explore intricate relationships

between variables without prior assumptions. Their advantageous properties make them well suited for describing non-linear, complex relationships, which are particularly common in biological "big data" from modern cancer research. The application of increasingly advanced machine learning algorithms in cancer research, particularly in biomarker research, could significantly improve the prognosis of cancer patients.

1.6 Research need, the subject of the thesis

Numerous studies have highlighted the role of EVs in tumorous processes, leading to efforts to include them in liquid biopsy based diagnostic methods25. 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).

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 central nervous system (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. A 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, brain metastasis, meningioma, and lumbar disc herniation, 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. 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. **Table 1** provides a summary of the materials and methods used in the meta-analysis.

Table 1. *Materials and methods of the meta-analysis.*

Material or method	Detail/purpose
Proteomic dataset	It contains the proteome of EVs isolated from the supernatants of NCI-60 cell lines (6,701 proteins). The intensities were used and logarithmized. NCI-60 cell line panel represents nine tumor types: breast, CNS, colon, kidney, leukemia, lung, melanoma, ovarian, prostate.
Invasion capacity data	It includes the invasion capacity of NCI-60 cells measured with CIM Plate-16.
Proliferation capacity data	Data on doubling time of NCI-60 cell lines.
RNA expression data	Microarray gene expression data of NCI-60 cell lines.
In situ tissue expression data	Information on protein expression in tissue and patient survival time.
Classification of EV samples	Multivariate logistic regression applied to classify EV samples into nine tumor types based on proteomic data. Feature selection performed using Least Absolute Shrinkage and Selection Operator (LASSO).
Regression for invasion and proliferation capacity	Multivariate linear regression with LASSO used to predict invasion and proliferation capacity of cell lines.
Pathway enrichment analysis	Conducted to identify overrepresented biological processes, molecular functions, and cellular components in proteomic data using Gene Ontology and Reactome.
Hierarchical clustering	Performed clustering based on proteins and Reactome results for visualization and analysis.
t-distributed Stochastic Neighbor Embedding	Utilized for 2-dimensional visualization of proteomic data.
Examining similarity between EV proteome and cellular RNA profile	Investigated similarity between EV and cellular RNA profiles using Spearman's correlation analysis and RV coefficients.

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

Table 2 provides a summary of the materials and methods used in out clinical study.

Table 2. *Materials and methods of the clinical study..*

Material or Method	Detail/Purpose
Patients	Blood samples from 138 patients were obtained from patients with glioblastoma (GBM), brain metastases (BM), and spinal disc herniation (CTRL) were obtained with consent and ethical approval.
Preparation of serum samples, sEV isolation and characterization	Small extracellular vesicles (sEVs) were isolated from blood samples via differential centrifugation and characterized using NanoSight NS300 for size distribution and Western blot for classical EV markers. Based on the characterization results, they were considered as sEV-enriched isolates and not as pure isolates.
Raman spectroscopy	Characterization of sEV-enriched isolates using Raman spectroscopy with a Senterra II Microscope.
Data adjustment	Baseline-corrected Raman spectral data were normalized using the Standard Normal Variate (SNV) method and subjected to Principal Component Analysis (PCA) for dimensionality reduction.
Classification	Utilized linear Support Vector Machine (SVM) algorithm to classify patient groups. Evaluated classification models for sensitivity, specificity, and AUC values.
Determining the spectral differences	Evaluated correlation between principal components (PCs) and patient groups using FreeViz method. Analyzed statistical differences between PCs to identify spectral variations.

4. NEW FINDINGS

The following chapter summarises the new findings described in detail in the PhD thesis (framed) and the related results.

---

1. Machine learning methods demonstrated that extracellular vesicles carry a highly tumor-specific molecular pattern.

Results related to the finding:

The proteomic dataset in our meta-analysis, which included 60 EV samples, uncovered a total of 6,071 proteins. Out of these, 5,908 proteins were quantified by intensity, collectively comprising what we refer to as the entire proteome in our study. Our analysis pinpointed 213 proteins consistently found across all EV samples, which we define as the core proteome. Initially, classification processes were carried out based on the core proteome using multivariable logistic regression, yielding an efficiency of 49.14%. This efficiency improved to 69.10% when the entire proteome was incorporated into the model. Subsequently, by employing the LASSO selection method, we compiled a discriminative protein panel from the entire proteome (172 proteins) and conducted classifications based on these, achieving a classification efficiency of 91.67%. The classification efficiencies obtained significantly exceeded the random classification efficiency of 11.11% for nine tumor types.

---

2. The proteomic content of extracellular vesicles may reflect tumor-specific signaling pathways.

Results related to the finding:

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.

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.

---

3. Protein panels compiled from extracellular vesicles' proteome can be used to estimate invasiveness and proliferative capacity of donor cells.

Results related to the finding:

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

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.

Results related to the finding:

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.

After averaging the spectra, row normalization was performed using the SNV method-

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).

Pairwise comparisons were conducted using the linear SVM 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.

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.

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. 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).

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

Our study leveraged a combination of machine learning algorithms to navigate the complex landscape of biological "big data". Specifically, we utilized linear and logistic regression coupled with LASSO, a method known for its efficacy in variable selection, and SVM algorithms. LASSO's capability to shrink less important variables to zero proved invaluable in distilling pertinent information from proteomic datasets, thereby facilitating the creation of interpretable and generalizable models. Similarly, SVM's robust handling of spectral data enabled accurate classification, particularly beneficial in scenarios where variables outnumber samples, a common occurrence in spectral analysis. Our approach not only identified tumor-specific protein patterns carried by EVs but also provided insights into invasion capacity, proliferation time, and tumor characteristic pathways, culminating in efficient classification of CNS tumor patients based on Raman spectra of small EV-enriched isolates from serum samples.

5.2. Raman spectroscopy of EVs as a promising diagnostic approach

Our research focused on harnessing Raman spectroscopy to unlock the diagnostic potential of EVs, particularly in the realm of CNS tumors. Unlike prior studies that primarily targeted limited number of biomarkers (only proteins, lipids or nucleic acids), our approach embraced the entire molecular profiling of sEV-enriched isolates, offering a more comprehensive diagnostic strategy. Our machine learning-based classification models, rated as "excellent" and "outstanding" based on ROC analysis standards, demonstrated exceptional accuracy in distinguishing between control groups and various CNS tumor types. Remarkably, our model's ability to discriminate between CTRL and GBM samples was particularly noteworthy, underscoring the method's efficacy in addressing the diagnostic challenges posed by highly variable tumors like GBM. Furthermore, our study filled a notable gap in the literature by pioneering the classification of CNS tumors based on Raman spectra of sEV-enriched isolates from serum samples.

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

Our investigation elucidated the remarkable tumor-specific molecular patterns carried by EVs, as evidenced by Raman spectroscopic analysis of clinical samples and meta-analysis of proteomic data. Previous studies have demonstrated the utility of EV analysis in distinguishing between cancerous and non-cancerous samples and categorizing different tumor types based on their molecular characteristics. Our meta-analysis aimed to further differentiate various tumors based on proteomic data, revealing a high degree of specificity in tumor-specific molecular signatures carried by EVs. 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.

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. 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

While the diagnostic potential of EVs has been extensively explored, their prognostic implications remain relatively uncharted area. Our study aimed to fill this gap by investigating the prognostic value of EV molecular content, particularly in estimating tumor cell invasion capacity and proliferation rates. Our meta-analysis revealed a subset of the EV proteome associated with these key factors in tumor progression and metastasis.

Our results revealed that certain proteins selected from the proteome of EVs have high predictive values for cell proliferation time and invasion capacity. Moreover, the tissue expression of these proteins is consistent with our results in influencing patient survival.

Considering 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 molecular characterization of tumors is crucial for effective treatment strategies, for example, aiding in the identification of suitable drug targets or develop personalized treatment strategies. This usually requires tissue biopsy, but has limitations due to invasiveness and intratumoral heterogeneity. Therefore, there is growing interest in liquid biopsy approaches.

Recent reviews indicate that tumor-derived EVs carry specific signaling and metabolic pathway components. Analyzing EV molecular content offers insights into tumor characteristics, microenvironment interactions, and distant communication.

Our meta-analysis has shown that the proteins that show the greatest differences between the nine tumour types are significantly associated with specific molecular mechanisms and signalling pathways. In many cases, the literature was in accordance with our results, as the correlation between different tumour types and different signalling pathways was experimentally demonstrated.

These results suggest EVs could serve as messengers of tumor strategies, aiding in drug target identification, personalized medicine development, and enhancing understanding of tumor biology.

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. 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.