BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

CENTRAL EUROPEAN INSTITUTE OF TECHNOLOGY BUT

STŘEDOEVROPSKÝ TECHNOLOGICKÝ INSTITUT VUT

DEVELOPMENT OF A DEVICE AND METHODOLOGY FOR LASER-INDUCED BREAKDOWN SPECTROSCOPY (LIBS)

VÝVOJ ZAŘÍZENÍ A METODIKY PRO SPEKTROMETRII LASEREM BUZENÉHO MIKROPLAZMATU (LIBS)

SHORT VERSION OF DOCTORAL THESIS

TEZE DIZERTAČNÍ PRÁCE

AUTHOR — Ing. Erik Képeš

AUTOR PRÁCE

SUPERVISOR — prof. Ing. Jozef Kaiser, Ph.D.

ŠKOLITEL

BRNO 2021

Abstract

This work deals with the transfer of analysis models across various laser-induced breakdown spectroscopy (LIBS) systems and the comparison of LIBS measurements obtained on distinct systems. Both the LIBS instrumentation and the data processing applied to LIBS data are highly flexible. Unfortunately, due to these flexibilities, results obtained on one LIBS system are rarely directly comparable to results obtained by a different system. This is further complicated by the various, often unknown, impact of the wide range of data processing algorithms on the LIBS data. Consequently, analysis models are generally system (and parameter) specific. The transfer of analysis models across various systems would result in the significant enhancement of the analytical capabilities of LIBS and in moderate cost reductions in industrial LIBS applications. The work studies the impact of various data collection strategies on LIBS data. Moreover, the work investigates the transformation of the acquired LIBS data through data processing. Lastly, the work addresses the transfer of analysis models across various LIBS systems.

Abstrakt

Táto práca sa zaoberá prenosom analytických modelov medzi rôznymi systémami spektroskopie laserom indukovanej plazmy (LIBS) a porovnaním LIBS výsledkov získaných na rôznych systémoch. Instrumentácia LIBS aj spracovanie LIBS spektier sú vysoko flexibilné. Bohužiaľ, kvôli týmto flexibilitám sú výsledky získané na jednom LIBS systéme zriedka priamo porovnateľné s výsledkami získanými na inom systéme. Toto je ďalej komplikované rôznymi, často neznámymi, účinkami algoritmov spracovania LIBS spektier. V dôsledku toho sú modely analýzy spravidla špecifické pre systém (a parametre). Prenos analytických modelov medzi rôznymi systémami by viedol k významnému zlepšeniu analytických schopností metódy LIBS a k miernemu zníženiu nákladov v priemyselných aplikáciách LIBS. Práca skúma vplyv rôznych stratégií merania metódou LIBS. Naďalej, práca skúma transformáciu získaných LIBS spektier prostredníctvom spracovávania údajov. Práca sa napokon zaoberá prenosom analytických modelov medzi rôznymi LIBS systémami.

Keywords

laser-induced breakdown spectroscopy, plasma characterization, laser ablation, machine learning, spectroscopic data processing

Klíčová slova

spektroskopie laserem indukovaného plazmatu, charakterizace plazmy, laserová ablace, strojové učení, zpracování spektroskopických dat

KÉPEŠ, E. *Development of a device and methodology for Laser-Induced Breakdown Spectroscopy (LIBS)*. Brno: Brno University of Technology, Central European Institute of Technology BUT, 2021. 51 p. Supervisor: Prof. Ing. Jozef KAISER, Ph.D.

Table of Contents

Introduction

Material identification—while often performed without conscious effort—is an integral part of our everyday lives. While opening a door, we can effortlessly distinguish between a metal and a plastic door handle and infer about the quality and durability of the handle. A more conscious effort is made towards material identification while shopping for new kitchenware. Namely, we would generally like to differentiate between cheap and possibly dangerous aluminium pots and high-quality stainless-steel ones. Going further, we might also attempt to tell the difference between an original branded piece and a knock-off made of a lower steel grade.

These tasks present increasingly more difficult challenges. Meanwhile, they are carried out millions of times around the world every day, e.g., in the context of quality assurance. Correspondingly, a variety of techniques have been developed. These techniques predominantly rely on spectrometry—the measurement of a physical property as a function of a controlled measurement parameter, e.g., laser-induced breakdown spectroscopy (LIBS), which uses a laser pulse to vaporize, excite, and ionize the target material [1]. Subsequently, the ablated material forms a laser-induced plasma (LIP). The plasma dynamically expands, interacts with the surrounding atmosphere, and cools down. The cooling process is accompanied with strong optical emission. The emitted radiation can be used to identify the atomic species present in the target. LIBS provides superior robustness, flexibility, and acquisition speed (sampling speed) [2, 3] compared to most other spectroscopic techniques. Specifically, LIBS can sample solid, liquid [4], and gaseous [5] materials without the need of extensive sample preparation. Considering the sampling mechanism, LIBS can be naturally used for depth profiling (3D resolution of the sample composition) [6, 7]. Moreover, the laser pulse can be focused to relatively large distances. Thus, LIBS is capable of remote (stand-off) analysis [8]. Additional considerable benefits of LIBS are the richness of the provided emission spectra, which span several spectral regions (from the ultraviolet, through the visible, to the near infrared) and the simultaneity of the spectral acquisition—that is, the whole spectral region can be detected from a single measurement. Consequently, LIBS is an often-preferred choice for many in-situ and laboratory analyses, e.g., of geological, biological, and industrial samples. Most notably, LIBS is part of the instrumental suite of both the Curiosity [9] and Perseverance [10] Mars rovers. Naturally, the benefits offered by LIBS come as trade-off for higher limits of detection, lower sensitivity, and low measurement repeatability compared to other AES techniques [2]. In addition, LIBS is often characterized with considerable spectral interferences and matrix effects [11].

The past few decades have been marked with the significant development of the LIBS instrumentation with the aim of mitigating the limitations of LIBS. In fact, the current limitations of LIBS mainly lay in the analysis of the acquired data. While considerable efforts have been made towards the development of signal processing methodologies that address spectral and matrix interferences and to improve the measurement repeatability, the full potential of LIBS is not yet in sight. Namely, LIBS is not yet capable of robust, repeatable absolute analysis, i.e., the determination of the sample's composition without relying on calibration standards. Thus, LIBS continues to rely on calibration models trained on extensive spectral libraries. However, the construction of robust spectral libraries can incur significant costs. As an example, the dataset used to calibrate the ChemCam LIBS instrument consists of over 400 specimens. This might appear extreme but addressing the various matrix effects and the non-linear LIBS signal response necessitates the coverage of a wide range of chemical compositions in distinct material matrices. Therefore, it is highly desirable to reuse existing spectral

libraries, both for quantitative and qualitative analyses. Nevertheless, our ability to do so is currently severely limited owing to the side effects of the main advantages of LIBS. LIBS is rather sensitive to the ablation pulse's parameters and the surrounding atmosphere, to name a few (a significant portion of Part I is dedicated to describing the parameters affecting the LIBS signal). Hence, seemingly minor changes in the measurement conditions can considerably alter the recorded signal. Moreover, the instrumental flexibility of LIBS means that two LIBS systems rarely possess the same spectral range, resolution, and spectral response. Consequently, the present thesis aims at describing the efforts made towards developing a methodology allowing the transfer of spectral libraries across distinct LIBS systems and measurement conditions.

The full version of this work comprises three parts. **Part I** describes the physical processes involved in the generation of the LIBS signal. The aims of Part I are twofold. Firstly, the understanding of the physical processes dictating the generation of the LIBS signal allows a more complete appreciation of the complexity of LIBS and its limitations. Secondly—and more importantly—reviewing the state-of-the-art (SOTA) understanding of the signal-generating processes and the impact of the various measurement and instrumental parameters allows the identification of the steps necessary for the transfer of spectral libraries.

Subsequently, **Part II** presents the current state of the library transfer methodology. Two approaches are considered: 1) the transformation of the collected spectra (spectral library) into a form compatible with the altered measurement conditions; and 2) the modification of the calibration models to address the changing measurement conditions. Thus, not only the data analysis techniques most frequently used for building calibration models (regression models for quantitative analysis and classification models for qualitative analysis) are presented, but also the various approaches to standardizing LIBS spectra.

Finally, **Part III** summarises and critically evaluates the author's contribution towards the improvement of the LIBS data analysis methodology. As such, Part III is centred around the author's peer reviewed publications thematically separated into six chapters, each addressing a different stage of the LIBS analysis, such as signal acquisition, spectra standardization, and the construction of calibration models.

1 Physical Principles of Laser-Induced Breakdown Spectroscopy

This chapter summarizes the first part of the full version of the dissertation thesis, which introduces the plasma properties that are necessary to understand the generation of LIBS spectra. Namely, the aspects of plasma properties that are the most crucial for describing the LIBS spectra are discussed. Subsequently, the various aspects of the optical emission of LIPs relevant to LIBS are presented. In addition, as the LIP is a highly transient phenomenon, the temporal evolution of LIPs is briefly presented. Then, the influence of the ablation process and the sample's properties on the plasma properties are reviewed. Lastly, several considerations regarding the collection geometry, spectral resolution, and the detection timing are discussed.

1.1 Thermodynamic Equilibria in Plasmas

A plasma is a gaseous quasi-neutral collection of neutral atoms, negatively charged electrons, and positively charged ions that exhibit a collective behaviour. In general, similarly to other gaseous media, plasmas consist of a large number of particles, e.g., typical number densities in LIBS plasmas range between 10^{16}—10^{19} cm^{-1}. In addition, the charged particles (electrons and ions) generate, interact with, and are affected by the electric and magnetic fields present in the plasma. Thus, the complete description of plasmas in terms of the position and velocity of every individual particle and the electric and magnetic fields (i.e., a complete kinetic description) is infeasible. Instead, plasmas are characterized statistically in terms of distribution functions and their statistical moments, which are not measured directly but inferred from the emission spectra. Specifically, the determination of the number densities N_a of species a—the zeroth (scalar) moment—from the emission line intensities present in an observed LIBS spectrum is the central problem of quantitative LIBS analysis.

For the statistical moments to be applicable, the plasma must be in (or close to) a local thermodynamic equilibrium (LTE). A plasma contains various forms of energy, namely kinetic, excitation, ionization, and radiative energy. These, in turn, are generally described by distinct distribution function, each characterized by an individual temperature value:

- The kinetic energy, which is dominated by the contribution of electrons, is described by the Maxwell—Boltzmann distribution and the kinetic temperature T_{kin}.
- The excitation state of the species, given by the relative population of neutral and various excited states in the atoms and ions is described by the Boltzmann distribution. The Boltzmann distribution is maintained *via* excitation and de-excitation processes and is characterized by the excitation temperature T_{exc}.
- The ionization state of the species is described by the Saha distribution and characterized by the ionization temperature T_{ion}.
- The radiative energy is described by the Planck distribution and characterized by the radiation temperature T_{rad}.

In LIBS, LTE conditions are generally considered. Under LTE conditions, the following relationship between the temperatures holds: $T_{kin} = T_{exc} = T_{ion} \neq T_{rad}$.

The importance of the presence of LTE conditions is twofold. Primarily, it is an underlying assumption of most methodologies developed and routinely applied to the characterization of LIPs, e.g., for the determination of the plasma temperature T and electron number density N_e. Secondly, the comparison of optical emission spectra obtained from different LIBS systems is significantly simpler, if only LTE plasmas are considered; LTE conditions allow the decoupling of the plasma characteristics from the instrumental parameters.

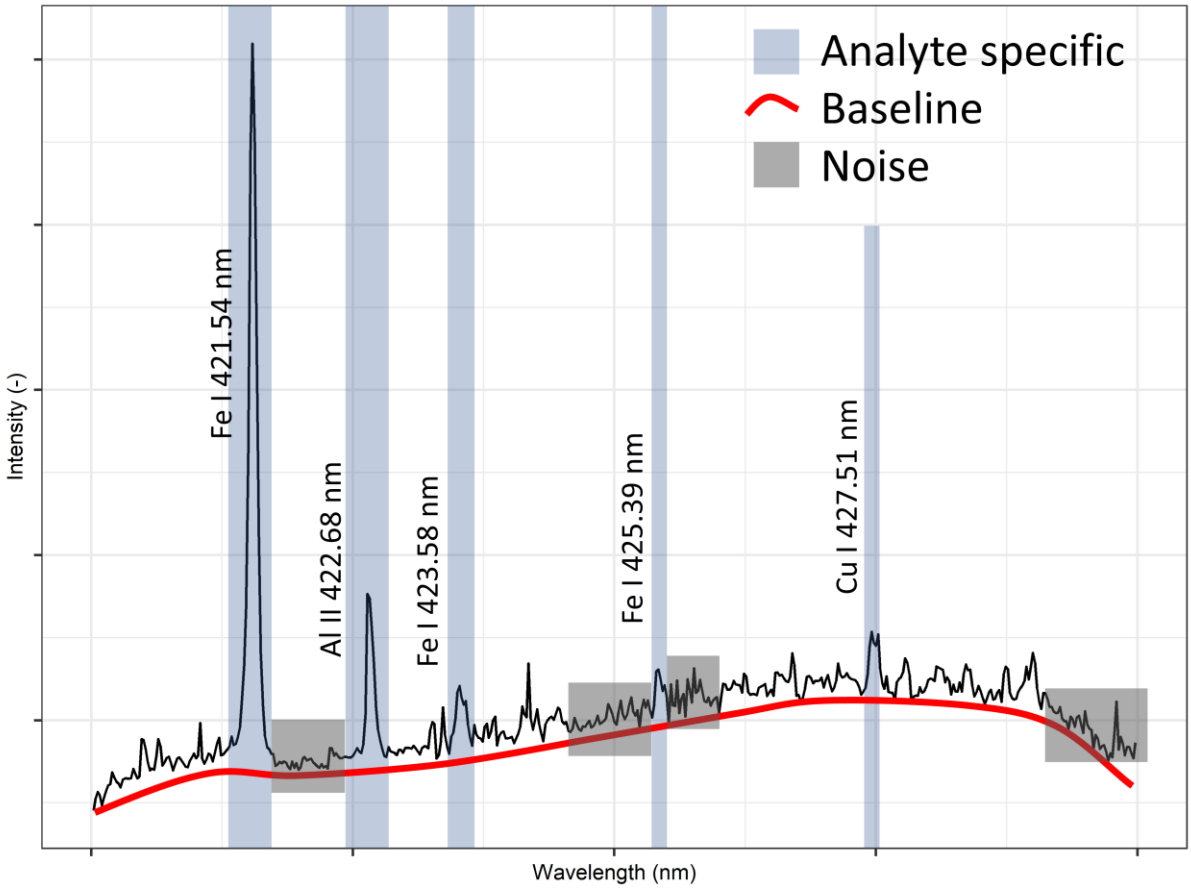

Figure 1-1 LIBS spectrum of the EB316 certified Al alloy standard with examples of the two signal components and noise highlighted. The spectrum was obtained using a single 80 mJ pulse with a wavelength of 1064 nm, a t_d of 400 ns, and a t_g of 50 µs.

1.2 Optical Emission and Absorption of Plasmas

Laser-induced breakdown spectra are composed of two major sources of optical signal: 1) Bremsstrahlung radiation and recombination radiation yielding a non-characteristic continuum radiation with limited analytical value; 2) the spontaneous emission of excited atoms or ions which yields characteristic and analytically valuable information. In addition, in some cases, such as the analysis of organic samples, various molecular emission bands can also be observed. The sum of these contributions yields an emission spectrum (Figure 1-1).

In the simple case of a completely optically thin plasma, the analyte-specific line emission at the photon frequency $\nu_{ul}^{(a)}$ corresponds to the electron transition in the species a from the upper energy state u to the lower energy state l with a corresponding energy difference E_u. The exact shape of the emission line is determined by the various line broadening mechanisms present in the plasma. Namely—in decreasing significance—pressure broadening (Stark effect), instrumental broadening,

Doppler broadening, and natural broadening. The natural and pressure broadening of the line profile can be approximated by a Lorentz. On the contrary, the Doppler and instrumental broadenings are best described by a Gaussian profile. Continuum radiation in the context of LIBS arises from two sources, the interaction of free particles resulting in Bremsstrahlung and the absorption of free electrons by ions resulting in recombination radiation.

If the LIP's absorption of optical radiation cannot be neglected, the LIP is referred to as optically thick. Consequently, the plasma can absorb radiation at specific frequencies, which is quantified by the absorption coefficient. The line emission coefficient and the absorption coefficient are related via Kirchhoff's law of radiation. Moreover, the optical thickness of the plasma can spatially vary. In this case, the optical thickness of the plasma is given by integrating the absorption coefficient along the line of sight. Apart from influencing the integrated spectral radiance, the optical absorption by the LIP (commonly referred to as self-absorption) also affects the perceived line shape.

1.3 The Laser-Induced Plasmas' Spatial and Temporal Inhomogeneities

One of the major challenges of LIBS is the temporal transiency of the emitting LIP [12]. Hence, understanding how the plasma evolves in time and space is critical. There are several simplified models available for describing the LIP's expansion, such as the Sedov—Taylor (ST) strong explosion model and the drag model, each applicable to a limited temporal window of the plasma's lifetime. For more detailed insights, state-by-state approaches must be used [13, 14], e.g., by solving the Euler equations or the Navier—Stokes equations. This, the temporal evolution of the LIP represents a chaotic system, where small changes in the initial conditions can result in considerable changes at the later stages of the LIP's lifetime. Consequently, even small changes in the measurement conditions can yield significantly different LIBS spectra.

The commonly observed spatial distribution of LIP properties as follows: the N_e and T tend to decrease along both the radial and vertical distances, while the plasma's centre was observed to be fairly homogeneous [15]. On the contrary, the maximum of the distribution of the atomic emissivity was reported to be radially shifted towards the plasma boundary [16]. Significant inhomogeneities of the temperature's and number densities' distributions were observed along the axial dimension as well. Namely, the highest atomic line intensities were observed about $1\ \mathrm{mm}$ above the sample surface, from where the intensity rapidly decreased in both directions [17].

In addition to the spatial inhomogeneities of LIBS plasmas, they are also temporally transient. Namely, LIPs expand into the ambient atmosphere. As the LIP expands and interacts with the atmosphere, the LIP obtains its characteristic shape. In addition, the positions and magnitudes of the various maxima described above slowly vary in time [18, 19]. Consequently, the ambient gas's pressure, molecular mass, and thermal conductivity have a pronounced influence on the LIP's properties.

1.4 Laser Ablation and Matrix Effects

The LIP's evolution is significantly affected by the initial conditions [20], i.e., the ablation and the closely related material properties of the specimen material (e.g., electrical and thermal conductivity, complex refractive index, melting and boiling points, and various mechanical properties). Namely, the LIP's characteristics are significantly affected by the material introduced into the plasma, i.e., by the ablated mass. Moreover, the ablated mass may not completely dissociate in the LIP, leading to particle formation. In turn, laser ablation is heavily affected by the matrix of the analysed material; the same material will produce different LIBS spectra when sampled in soil or pellet form.

In addition, laser-induced ablation and the evolution of the corresponding LIP is linked to the ablation laser pulse's properties. For example, the ablation mechanisms change when ns and fs pulse lengths are considered [21]. Thus, it is challenging to draw brief and comprehensive conclusions regarding the influence of the laser parameters. Nevertheless, the pulse duration, pulse energy, pulse shape, and laser wavelength were shown to affect both the signal intensity and its stability by yielding plasmas with different temperatures, electron number densities, degrees of ionization and ablating varying amounts of target material.

One of the major assumptions of LIBS analysis is the achievement of stoichiometric ablation, i.e., the plasma's atomic composition is equal to that of the analysed sample. Despite achieving LIP formation, the stoichiometry of the produced plasma might not be ensured. A major concern for the fulfilment of the stoichiometric conditions is fractionation. Namely, the species contained in the sample material generally possess distinct melting and evaporation temperatures. Nevertheless, fractionation is mainly limited to copper-based alloys, such as brass and bronze [22]. Apart from fractionation, several other matrix effects are generally considered, such as line and elemental interferences, sample moisture, sample temperature, sample matrix, surface roughness, and grain size. Collectively, matrix effects distort the expected (ideal) relationship between the sample content and the analytical signal, i.e., the curve-of-growth (COG) [23].

1.5 Collection, Resolution, and Detection

There are several additional factors affecting the LIBS signal. Namely, the components used to collect, transfer, and resolve the LIP's optical emission—collectively referred to as the optical train—and the detection camera [24]. There are four general considerations regarding the optical train and the camera in the context of the transfer of calibration models:

1) The optical train's ability to collect light from a selected region of the LIP is necessitated by the plasma's spatial inhomogeneities. Namely, small collection angles allow the observation of a narrow, homogeneous part of the plasma. In contrast, large collection solid angles are less sensitive to the fluctuations of the plasma's location, but the collected emission originates from various plasma regions which can be characterized by distinct T and number densities.

2) The spectral resolution of the optical train, defined in terms of its resolving power. In general, a system with a higher resolving power might detect emission lines that are not present in the spectra obtained with a lower resolving power LIBS system. Nevertheless, this could be addressed by downscaling the spectra obtained with a higher resolving power.

3) The spectral range of the optical train, which is largely defined by the type of monochromator used. Differences in spectral range could be addressed by truncating the spectra obtained with a spectrometer exhibiting a wider spectral range.

4) Considering the temporal transiency of LIPs, the temporal gating of the emission's detection should be considered. Namely, the existence of an optimal delay time t_d [25] and gate-time t_w [26] have been reported. However, their relation to the plasma's parameters is unclear.

2 The State of Library Transfer in Laser-Induced Breakdown Spectroscopy

Part II of the full version of this thesis, which is summarized in the following chapter, provides an overview of how model transfer could be attempted and the current state-of-the-art in LIBS. As such, it gives a brief overview of transfer learning (TL) and re-defines library transfer as a transfer learning problem. Thus, LIBS spectra are regarded as statistical observations and treated as vectors in a high-dimensional space. Subsequently, the following sections detail the current SOTA of the LIBS data' processing and its relevance to transfer learning. Namely, the various existing approaches to spectral standardization are reviewed and several calibration models which might could be potentially applied for transfer learning are presented. Namely, partial least-squares (PLS), decision trees (DTs), support vector machines (SVMs), and artificial neural networks (ANNs).

2.1 Spectroscopic Data

In the context of data analysis, LIBS spectra are high-dimensional data points $\mathbf{x} \in \mathbb{R}^Q$. Thus, a spectral library represents a data matrix $\mathbf{X} \in \mathbb{R}^{M \times Q}$, where M is the number of spectra and Q corresponds to the number of covariates (wavelengths). In addition, a matrix $\mathbf{Y} \in \mathbb{R}^{M \times R}$ contains the dependent variables. Namely, LIBS spectra exhibit a high degree of:

- Sparsity – A significant portion of the covariates carries no valuable information.
- Redundancy – Spectral features (emission lines) consist of multiple covariates, which carry the same (or very similar) information. In addition, the same species is often represented by multiple emission lines.
- Non-linearity – The LIBS signal's response is non-linear.

Moreover, LIBS is generally characterized by significant shot-to-shot variability [27] and in some applications a considerable noise [28] (especially in single-shot analyses). Consequently, LIBS spectra consist of three components: 1) a low frequency background signal; 2) mid-frequency analyte-specific emission lines (and bands in the case of molecular emission); and 3) high-frequency noise.

2.2 Calibration Models

There are two fundamentally distinct approaches to building calibration models: physics-based modelling [29] and data-based modelling [30]. Data-based models are generally more successful in addressing the impact of various non-linear effects [29] at the expense of being more susceptible to overtraining and more challenging to interpret. Regardless of the modelling approach used to build the calibration model, the model can be written as a prediction function $f(\cdot)$ which is generally parametrized with a set of parameters $\boldsymbol{\theta}$**.** The values of $\boldsymbol{\theta}$ are unknown and are estimated using a subset of $\mathbf{X}$ referred to as the training data $\mathbf{X}_{train}$ to obtain $\widehat{\boldsymbol{\theta}}$. Consequently, a calibration model can be given by the prediction function $f(\widehat{\boldsymbol{\theta}}, \mathbf{X}_{train})$. As such, changes in $P(\mathbf{X})$ will directly affect the prediction function as well. Thus, since $P(\mathbf{X})$ depends on the measurement conditions, the calibration model will be system specific [31, 32].

2.3 Transfer Learning

The system specificity of calibration models is generally addressed by transfer learning [33]. Namely, transfer learning studies knowledge transfer between the source domain $\mathcal{D}_S = \{\mathcal{X}_S, P_S(\mathbf{X})\}$ and the target domain $\mathcal{D}_T = \{\mathcal{X}_T, P_T(\mathbf{X})\}$, where $\mathcal{X}$ are the feature spaces and $P(\mathbf{X})$ are the marginal distributions of the covariates (Figure 2-1). Since LIBS spectra can be truncated and downsampled, generally $\mathcal{X}_S = \mathcal{X}_T$. Moreover, the task defined by $\mathcal{T} = \{\mathcal{Y}, f(\boldsymbol{\theta}, \mathbf{X})\}$, i.e., by a set of labels represented in the feature space $\mathcal{Y}$ and a prediction function $f(\boldsymbol{\theta}, \mathbf{X})$ must be considered.

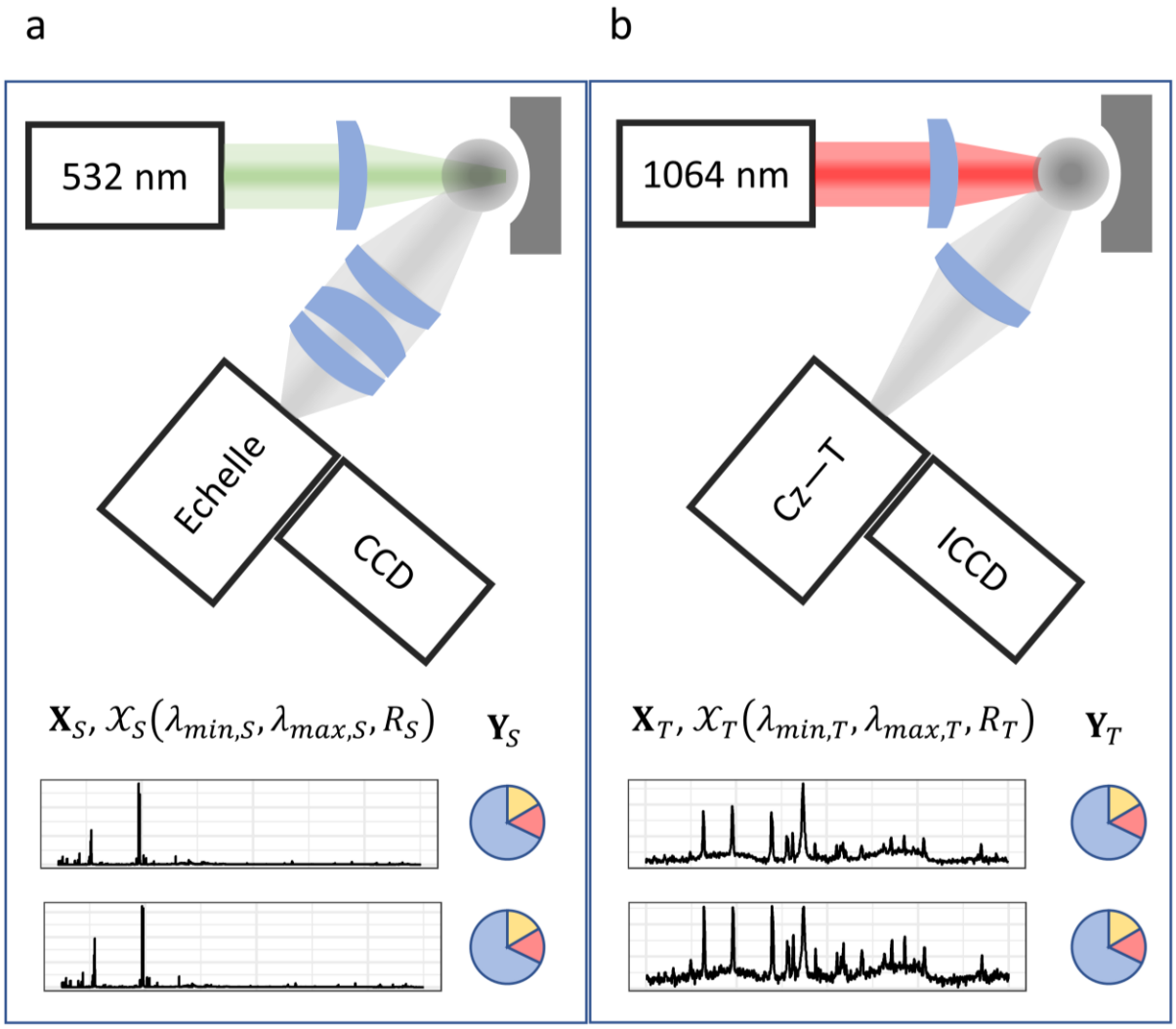

Figure 2-1 Summary of the transfer-learning problem. $\boldsymbol{X}_i$ denote the collected datasets (specifically, only the independent variables) with $i \in \{S — source, T — target\}$; $\mathcal{X}_i$ denotes the feature space of the dataset $\boldsymbol{X}_i$, i.e., the spectral range (from $\lambda_{min,i}$ *to* $\lambda_{max,i}$) and spectral resolution R_i used to collect $\boldsymbol{X}_i$; and $\boldsymbol{Y}_i$ denotes the feature space of the dependent variables (e.g., class membership for classification problems or a continuous material propery for regression problems). The spectra are single-pulse spectra of the EB316 certified Al alloy, collected with an ablation pulse wavelength of 1064 nm, pulse energy of 80 mJ, $t_d = 2$ µs, and $t_g = 50$ µs. $\lambda_{min,S} = 250$ nm, $\lambda_{max,S} = 900$ nm, $\lambda_{min,T} = 350$ nm, $\lambda_{min,T} = 370$ nm, $R_S = R_T = 6000$.

2.3.1 Data-based Transfer Learning

One approach to data-based transfer learning entails finding a common representation of the source dataset $\mathbf{X}_S$ and the target dataset $\mathbf{X}_T$ in which $\mathbf{X}_S$ and $\mathbf{X}_T$ exhibit a high level of correspondence. This is routinely performed in LIBS analyses on a small scale to minimize signal variations on a single system. These approaches include baseline correction (the removal of the low-frequency signal from the LIBS

spectra), denoising (the removal of the high-frequency part of the spectra), normalization (the scaling of the spectra to), and compensation for plasma properties (the correction of the spectra to a given temperature and number densities considering the theoretical emission of the plasma). Nevertheless, these approaches have not yet been demonstrated to achieve the intra-system standardization of LIBS data. Alternatively, assuming that an extensive labelled dataset $\mathbf{X}_S$ is available in $\mathcal{D}_S$, TL can be attempted by:

- Instance-based TL—If a limited number of labelled data points $\mathbf{X}'_T$ is available in $\mathcal{D}_T$, classifiers can be trained using subsets of $\mathbf{X}_S$ and $\mathbf{X}'_T$. $\mathbf{X}'_T$ is separated into training and validation sets. The subset of $\mathbf{X}_S$ is determined in an iterative process which assigns a higher weight to the data points of $\mathbf{X}_S$ which positively contribute to the correct classification of the validation subset of $\mathbf{X}'_T$ or reduce the quantification error on $\mathbf{X}'_T$ [34].
- Feature-based TL—While $P(\mathbf{X}_S)$ and $P(\mathbf{X}_T)$ generally differ, a subset of features might exhibit a similar probability density distribution in both $\mathcal{D}_S$ and $\mathcal{D}_T$. This subset of features can be found using an iterative approach while validating on the labelled portion $\mathbf{X}'_T$ of $\mathbf{X}_T$. Consequently, a calibration model trained on $\mathbf{X}_S$ using only an appropriate subset of features will perform well on $\mathbf{X}_T$ as well [35, 36].
- Representation-based TL—A common step in LIBS data processing is dimensionality reduction, which finds a low-dimensional representation of the initial dataset. This is commonly performed by finding an appropriate new set of basis vectors. In general, the basis vectors are selected, e.g., to maximize inter-class distances. However, using $\mathbf{X}'_T$ and $\mathbf{X}'_S$ (the labelled subset of $\mathbf{X}_T$ and the corresponding spectra from $\mathbf{X}_S$), an appropriate transformation can be found which that projects the observations from the same specimen to the same region of the embedding feature space [37, 38].

2.3.2 Model-based Transfer Learning

A common representation of $\mathbf{X}_S$ and $\mathbf{X}_T$ might not always be attainable. Thus, TL can be attempted considering the trained calibration models rather than the datasets. More specifically, existing models can be altered to incorporate the changes in measurement conditions, according to the following approaches:

- Fine tuning—The parameter set of a model $f(\widehat{\boldsymbol{\theta}})$ trained on $\mathbf{X}_S$ can be written as $\widehat{\boldsymbol{\theta}}_S = \widehat{\boldsymbol{\theta}}_0 + \widehat{\boldsymbol{\theta}}'_S$. Similarly, if the model is trained on $\mathbf{X}_T$, its parameters are $\widehat{\boldsymbol{\theta}}_T = \widehat{\boldsymbol{\theta}}_0 + \widehat{\boldsymbol{\theta}}'_T$. Here, $\widehat{\boldsymbol{\theta}}_0$ is common among the model trained on data in $\mathcal{D}_S$ and $\mathcal{D}_T$. Meanwhile, $\widehat{\boldsymbol{\theta}}'_S$ and $\widehat{\boldsymbol{\theta}}'_T$ are domain specific. Assuming that $\widehat{\boldsymbol{\theta}}_0$ dominates both $\widehat{\boldsymbol{\theta}}_S$ and $\widehat{\boldsymbol{\theta}}_T$, a well-performing calibration model can be trained on $\mathbf{X}_S$. Subsequently, the trained model is fine-tuned, i.e., it is trained further on a limited available labelled subset $\mathbf{X}'_T$ of $\mathbf{X}_T$ [39, 40].
- Data transformation—Some machine learning approaches—e.g., artificial neural networks—can approximate any function. Thus, an ANN can be trained to predict the shape of data points in $\mathcal{D}_S$ from unlabelled data points in $\mathcal{D}_T$ from a limited number of labelled observations $\mathbf{X}'_T$ in $\mathcal{D}_T$. Consequently, newly acquired LIBS spectra could be transformed using a trained ANN to match a selected domain [41].
- Model weighting—If several models are trained in $\mathcal{D}_S$, the models can be weighted according to their performance on the limited labelled data $\mathbf{X}'_T$ from $\mathcal{D}_T$. That is, an ensemble model (a combination of several models) can be trained that aggregates the output of several models and makes a final decision. Alternatively, if several source domains $\mathcal{D}_{S,i}$ are available, a single model can be trained in each $\mathcal{D}_{S,i}$, followed by the aggregation of the results as above [42, 43].

2.3.3 Transfer Learning in Spectroscopy

In the context of spectroscopy, transfer learning has been explored only to a limited degree, with most of the progress limited to applications in near-infrared (NIR) spectroscopy and LIBS. To a lesser extent, TL has been demonstrated in reflectance spectroscopy, mid-infrared (MIR) spectroscopy, and Raman spectroscopy. However, most applications addressed TL between domains that exhibit significantly less variability compared to the transfer of spectral libraries discussed in the present thesis. In what follows, a brief review of TL in spectroscopies is provided. Note that most of the current approaches predominantly use instance- and feature-based approaches and fine-tuning.

The simplest case of successful TL in LIBS was the simple inclusion of a few specimens from $\mathcal{D}_T$ into the training dataset with $\mathbf{X}_{T,trn}$. This improved the LIBS analytical performance of steel specimens with a rough surface finish when using a calibration model trained on smooth specimens [44].

Feature-based TL improved the quantitative performance of LIBS in the analysis of hot (above a thousand °C) steel samples [45]. A regression model was trained on room temperature data. Then a limited number of specimens were measured at the high temperature. The latter were used to identify an appropriate set of covariates that allowed the accurate quantitative analysis of the high-temperature specimens. In an alternative approach, the optimal set of covariates was found by evaluating the variable importance in projection of a partial least squares model [46]. In a different approach, only features that correlated with the predicted quantity were kept, which improved the calibration model's performance in $\mathcal{D}_T$ [47].

Finally, the transformation of the datasets $\mathbf{X}_S$ and $\mathbf{X}_T$ has been explored in LIBS: a transformation-based approach was applied to improve the quantitative analysis of chromium in high-temperature steel samples. Only the composition of the room-temperature calibration dataset was known. A kernel function was used to project the datasets $\mathbf{X}_S$ and $\mathbf{X}_T$ into a high-dimensional feature space while minimizing the discrepancy between the distribution of the labelled data in $\mathcal{D}_T$ and the corresponding data in $\mathcal{D}_S$ [48].

2.4 Dimensionality Reduction

The high dimensionality of LIBS spectra combined with the high redundancy and significant covariance exhibited by the covariates are detrimental to the performance of most multivariate analyses [49]. Thus, it is desirable to transform the LIBS spectra into a lower-dimensional embedding, i.e., to perform dimensionality reduction (DR). Choosing an appropriate DR technique can reduce the shot-to-shot variations by discarding noisy covariates. Most DR techniques can be reduced to the same fundamental steps [50]: 1) determine a similarity matrix between the observations $\mathbf{x}_m$ in $\mathcal{X}$ (the initial feature space); 2) find a new representation $\mathcal{S}$ with a lower dimensionality compared to that of $\mathcal{X}$, while maintaining the similarity matrix to the highest possible degree. In other words, DR is a mapping $f: \mathcal{X} \rightarrow \mathcal{S}$, or $f(\mathbf{X}) = \mathbf{S}$. Depending on the similarity metric used, DR techniques can be linear or non-linear.

2.4.1 Linear Dimensionality Reduction

By far the most common approach is to carry out principal component analysis (PCA) [51], which uses spectral covariance as the similarity metric. Consequently, the basis vectors of $\mathcal{S}_{PCA}$ are the eigenvectors $\mathbf{W}$ of the covariance matrix of $\mathbf{X}^T\mathbf{X}$ [52]. The obtained low-dimensional representation $\mathbf{S}_{PCA}$ is the projection onto the feature space defined by the eigenvectors $\mathbf{W}$, i.e., $\mathbf{S} = \mathbf{XW}$. Dimensionality reduction using PCA is achieved by discarding the n least important principal

directions, i.e., the truncating of the columns of $\mathbf{W}$. The importance of a principal direction is determined by the magnitude of the corresponding eigenvalue.

2.4.2 Non-linear Dimensionality Reduction

The application of non-linear DR techniques is currently limited in the LIBS literature [53, 54]. The first major difference between PCA and non-linear DR approaches is the chosen similarity metric. Namely, the similarity matrix $\mathbf{D} \in \mathbb{R}^{M \times M}$ containing the pairwise similarities of the observations contained in $\mathbf{X} \in \mathbb{R}^{M \times Q}$ can be expressed by a polynomial kernel function [55, 56], Gaussian function [57, 58], cosine, or even the Euclidean distance [59, 60] (which is a linear (dis)similarity metric). The similarity matrix is then regarded as a graph, with the nodes being the observations and the length of the edges being the pairwise similarities [61]. The graph allows the determination of the k nearest neighbours of every observation. Considering whether the full graph is used for the subsequent DR or only the local neighbourhood of the observations, DR techniques are classified either as full or sparse, respectively [50]. The last step is to find a low-dimensional representation $\mathcal{S}$, in which the similarity matrix is at least locally maintained. This frequently involves solving an eigenvalue problem (similarly to PCA), or iteratively with the aim of optimizing a loss function [62].

2.5 Common Aspects of Data-based Calibration Models

Data-based calibration models play a central role in transfer-learning as they can discover relationships hidden in complex datasets without the need of explicitly encoding these relationships. This is especially beneficial if highly non-linear relationships are suspected with an unknown shape, such is frequently the case in LIBS. Hence, instead of explicit programming, the relationships are found by constructing extensive datasets that cover the expected field of application.

The training of data-based models refers to the estimation of the model parameters $\widehat{\boldsymbol{\theta}}$, which is generally done by iteratively optimizing the so-called loss function $\mathcal{L}(\cdot)$[1] [69]. The exact form of $\mathcal{L}(\cdot)$ depends on the calibration model. However, an important general distinction can be made between models trained in a supervised or unsupervised manner. Namely, the loss function of regression and classification models that are trained in a supervise manner takes the general form of $\mathcal{L}(\mathbf{y}_i', \mathbf{y}_i)$, where $\mathbf{y}_i' = f(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}, \mathbf{X}_{trn})$ is a vector encoding the predicted dependent variables' values for the i^{th} observation and $\mathbf{y}_i$ is the corresponding ground truth value. Thus, $\mathcal{L}(\mathbf{y}_i', \mathbf{y}_i)$ compares $\mathbf{y}_i'$ and $\mathbf{y}_i$ and returns a scalar value. The training of the model is then carried out by altering $\widehat{\boldsymbol{\theta}}$ to lower the discrepancy between $\mathbf{y}_i'$ and $\mathbf{y}_i$. On the contrary, while training models in an unsupervised manner, $\mathbf{y}_i$ are not available.

2.6 Overtraining and Regularization

Building calibration models on datasets which have more covariates than observations leads to ill-posed optimization problems [70]. Thus, owing to the large number of covariates present in LIBS spectra, it is relatively straightforward to overtrain most ML models. An overtrained model performs well on the training dataset (in fact, it can often perform perfectly on the training data) but generalizes poorly on new observations, i.e., when the model is employed in a practical application. This is especially important if transfer-learning is considered; reducing the number of covariates reduces the difference between $P_S(\mathbf{X})$ and $P_T(\mathbf{X})$. There are two common strategies that can suppress the risk of overtraining: regularization and rigorous cross-validation.

[1] Also often referred to as cost function.

Given enough trainable parameters (degrees of freedom), most ML techniques will find a decision function $f(\cdot)$ that perfectly fits the training dataset $\mathbf{X}_{trn}$. This can be alleviated by modifying the loss function $\mathcal{L}(\cdot)$ by adding a term that penalizes the use of too many parameters [71]. The most common modifications of the loss function are:

- The L1 norm—adding the sum of the parameters' absolute value. The L1 norm enforces an arbitrary degree of sparsity among the parameters, that is, an arbitrary number of parameters can be forced to be 0.
- The L2 norm—adding the square root of the sum of the squared parameters. The L2 norm limits the magnitude of every parameter and forces the weights of correlated independent variables to be similar.

Another regularization strategy takes advantage of the prediction function's dependency on the training dataset $\mathbf{X}_{trn} \subsetneq \mathbf{X}$ ($\mathbf{X}_{trn}$ is a proper subset of the whole dataset $\mathbf{X}$). Consequently, the training of the calibration model can be repeated with different randomly sampled $\mathbf{X}_{trn}$ datasets and evaluated on the $\mathbf{X}_{vld} = \mathbf{X} - \mathbf{X}_{trn}$ dataset. This procedure is referred to as k-fold cross validation (kCV) if it is repeated k-times [72]. Alternatively, in each iteration $\mathbf{X}_{trn}$ can be obtained by excluding the observations belonging to a selected class. Naturally, $\mathbf{X}_{vld}$ will then consist of the observations of a single class. This process is referred to as leave-one-out CV (LOO-CV) [73].

2.7 Model Evaluation

To determine whether the calibration model transfer was successful, the trained model's performance must be evaluated on a test dataset. The evaluation is performed according to performance metrics relevant for the respective application.

2.7.1 Clustering

Clustering is an unsupervised task. Thus, the evaluation of clustering algorithms is not straightforward. Nevertheless, two options are available, depending on whether the true class memberships (ground truths) of the clustered data points are available or not.

- Supervised evaluation—If the class memberships are available, the Rand score [74].
- Unsupervised evaluation—If the class memberships are not available (which is commonly the case), various clustering indices can be used. For example, the silhouette score [75] enumerates how close the observations are to their assigned cluster's centre compared to the centres of the other clusters. An alternative clustering index is the Davies—Bouldin index [76], which relates the within-cluster scatter of the observations to the inter-cluster distances.

2.7.2 Classification

Classification can be broadly regarded as the supervised extension of clustering. Thus, the class memberships—which are encoded with discrete values—are expected to be known. Consequently, the class membership predicted by the classification model can be compared to the true class membership. Considering a binary classification or one-against-all classification scenario, there are four different outcomes of the comparison: true positive, i.e., the observation is correctly assigned to the reference class; true negative, i.e., the observation is correctly classified as not belonging to the reference class; false positive, i.e., the observation is incorrectly assigned to the reference class; and false negative, i.e., the observation is incorrectly classified as not belonging to the reference class. According to these four outcomes, several performance metrics are available for each class, e.g., accuracy, specificity or true negative rate, sensitivity or true positive rate, and false positive rate. Since

these metrics are determined for each individual class, in a multiclass classification task, they must be aggregated to determine the model's performance. This is commonly done by taking the average of the class-specific values.

2.7.3 Quantification

The aim of quantitative calibration models (regression models) is to predict a continuous value from the input vector. As such, the performance metrics used to evaluate quantitative models differ from those use for classification. Some of the most common quantitative figures of merit are the root mean squared error and the mean absolute error.

2.7.4 Pre-Processing and Dimensionality Reduction

While spectral pre-processing and dimensionality reduction are a critical part of data analysis, the evaluation of pre-processing techniques is challenging, especially on real data (as compared to synthetic data). Consequently, pre-processing is most often evaluated indirectly. That is, rather than defining a performance metric directly for the various pre-processing approaches, classification and regression are performed on datasets pre-processed in different ways. Consequently, the different pre-processing strategies are evaluated in terms of the classification and/or quantification performance of the calibration model(s) trained on the embedded data.

2.8 Hyperparameter Tuning of ML Models

Most models have numerous hyperparameters, which must be tuned for optimal model performance. Some examples include the number of latent variables used by PLS, the kernel function for SVMs, the number of nodes in DTs, and the network architecture in ANNs. In contrast with model parameters, the hyperparameters are optimized with regards to a human-interpretable performance metric rather than with respect to the loss function's value. These performance metrics are not differentiable with respect to the hyperparameters. Thus, the models must be retrained and evaluated for every considered combination of hyperparameters. This can be performed according to the following approaches:

- Grid-search—The comparison of every combination of the hyperparameters' values.
- Randomized grid Search—The comparison of randomly selected combinations of the hyperparameters.
- Genetic algorithms—The comparison of models with random hyperparameters followed by the selection of the best-performing models.

2.9 Selected Calibration Models

There are four calibration models, whose applicability is well-established in the LIBS literature, and which have been considered for transfer learning in general: partial least squares (PLS), decision trees (DTs), support vector machines (SVMs), and artificial neural networks (ANNs). Each model should be considered owing to distinct criteria and characteristics.

2.9.1 Partial Least-Squares Models

Dimensionality reduction can make or break a calibration model. In addition, finding an appropriate representation of the data is also a central challenge for transfer learning. Nevertheless, dimensionality reduction is commonly done in an unsupervised manner. However, if the response variables' values are available, they can be used to guide the dimensionality reduction technique towards a mapping which emphasizes covariates which are more important for the accurate

prediction of the response variables. One of the first methods developed to this aim is PLS, which is often viewed as the supervised extension of PCA. When applied to regression tasks, PLS is referred to as PLS regression (PLSR). Similarly, PLS discriminant analysis (PLS-DA) can be applied for classification tasks.

Apart from their popularity, PLS models should be considered for transfer learning owing to several additional reasons. Firstly, PLS models—unlike most ML techniques—can be easily interpreted. Thus, PLS models could perform favourably when combined with feature weighting approaches. In addition, regularized PLS models have been already used to identify and study matrix effects in LIBS [77]. Secondly, PLS models create a supervised feature representation. Thus, a limited amount of training data from the target domain $\mathcal{D}_T$ could be used to obtain feature representations that are appropriate for both $\mathcal{D}_S$ and $\mathcal{D}_T$. Nevertheless, PLS models are linear. Consequently, they cannot address the non-linearity of the LIBS response in a robust manner.

2.9.2 Decision Trees

Conceptually, decision trees (DTs) represent the simplest ML technique owing to their similarity to the common graphical representation of decision-making processes. Namely, DTs are constructed of a chain of nodes. The first (topmost) node is referred to as the root node, which represents a single input observation. Each subsequent node either makes a binary decision (considering a single feature from the feature space in which the data point is represented) about the observation being processed or terminates. In the case of the former, the node is referred to as a parent node and the decision creates two child nodes. If the node terminates, it is referred to as a leaf node and represents the output of the DT. An alternative representation of a DT is the iterative partitioning of a feature space in which the data points are represented.

Since the leaf nodes' values are determined relatively easily, the training of a DT refers to searching for the optimal decision rules. This is done by optimizing the feature and its value used for making the binary decision in each node with regards to a chosen loss function. Child nodes can be added until a termination condition is reached. Common termination conditions include limiting the total number of nodes (by choosing a maximum) or the change in the loss function's value (by choosing a minimum).

Decision trees should be considered for transfer learning owing to their decision-based formulation and straightforward interpretability [78]. As a consequence of the former, DTs could mimic the decision-making of a domain expert. The more straightforward interpretability of DTs [79, 80]—compared, e.g., with artificial neural networks—makes the troubleshooting of DTs, and hence, their training simpler.

2.9.3 Support Vector Machines

A classification problem can be regarded as finding a boundary line (or, in the general case, a hyperplane) that separates the observations belonging to the different classes. Similarly, regression problems can be viewed as the search for an appropriate curve that follows the observed data arbitrarily closely. This led to the development of various least-squares methods, and the more powerful support vector machines (SVMs). The latter can be used for both classification and regression.

In their most basic form, support vector classifiers (SVCs) carry out the binary classification of linearly separable data points [81]. That is, each input is assigned to one of two classes (labelled as $\{-1, +1\}$). The decision is made by projecting the individual data points onto the normal vector of the hyperplane that separates the data points representing two distinct classes $\mathcal{A}$ and $\mathcal{B}$ [82]. Multiclass classification

problems are reduced to a series of one-versus-all or one-versus-one binary classification problems where the final class membership is determined by an appropriate voting scheme [83]. Qualitatively, the hyperplane aims at maximizing its distance from the data points, i.e., the margin. In turn, the margin is defined by the data points closest to the hyperplane, which are referred to as support vectors.

The geometric idea of SVCs can be easily extended to regression tasks. Instead of finding the line that exhibits the smallest mean squared distance from every observation (as done during least-squares fitting), the support vector regression (SVR) algorithm finds a line with the smallest possible margin that includes all the observations. To avoid outliers influencing the regression line, a relaxation term is introduced which allows outliers to lay beyond the margin.

Support vector machines should be considered for transfer learning for two reasons. Firstly, the powerful geometric idea driving SVCs makes them considerably more robust against small to moderate shifts of $P(\mathbf{X})$ than other ML models: the data points can be freely translated as long as they remain separated by the trained hyperplane. Secondly, both SVCs and SVRs can employ the so-called kernel trick [84]. The kernel trick is essentially the projection of linearly non-separable data into a high-dimensional feature space where the data become linearly separable. This projection is done *via* a kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_{j\neq i}) = \varphi_i(\mathbf{x}_i) \cdot \varphi_j(\mathbf{x}_j)$, which creates a pairwise similarity/distance matrix of the dataset $\mathbf{X}$. Some common kernel functions are the Gaussian and polynomial kernels. The trick in using kernel transformations is that the transformation does not have to be explicitly carried out. Instead, the optimization problems underlying the training of SVM model are modified to incorporate the kernel function. Consequently, linear models can be fitted to non-linear data. Considering transfer learning, the construction of an appropriate kernel function can be attempted that addresses the shift in $P(\mathbf{X})$ [85].

2.9.4 Artificial Neural Networks

Most of the recent advances in numerous fields—such as computer vision, natural language processing, and even protein folding prediction—has been fuelled by artificial neural networks (ANNs). This class of machine learning models is attributed with the property of universal function approximation. Hence, in theory, ANNs are capable of approximating and parametrizing the convoluted processes generating LIBS data. This makes ANNs an attractive choice for transfer learning. Nevertheless, the current extent of the spectroscopic applications of ANNs is very limited.

The fundamental building block of both ANN architectures presented here is the artificial neuron (also sometimes referred to as a node). An artificial neuron is a simple mathematical operator, that takes a set of inputs and yields an output. More specifically, artificial neurons carry out the weighted summation of the outputs of the neurons in the preceding layer. The summation is followed by the application of a non-linear activation function $\sigma(\cdot)$. There are two distinct ANN architectures commonly applied for LIBS applications:

- Fully Connected ANNs (FCNNs)—In a FCNN, each neuron is connected to every neuron in both adjacent layers. That is, each neuron in the layer l sums up the (scalar) output of every neuron in the preceding layer $(l-1)$ and applied a non-linear activation function $\sigma(\cdot)$ to the weighted sum
- Convolutional ANNs (CNNs)—Convolutional ANNs differ from FCNNs in that CNNs contain at least a single convolutional layer. A convolutional layer comprises convolutional neurons, which apply the discrete convolution operation to the input vector (or matrix). Thus, in the case of a LIBS spectrum—which is represented by a 1D vector—a convolutional neuron will yield a transformed 1D vector. Historically, CNNs have been developed to address the spatial

invariance of image data. However, in LIBS, CNNs, are useful as they tend to consider whole emission lines as features rather than individual intensity values.

3 The Author's Contributions to the State of the Art

The contributions of the presented thesis to the development of the LIBS methodology are twofold. Firstly, efforts have been made to obtain a more complete understanding of LIBS plasmas, namely that of plasmas with a perturbed axial symmetry. Secondly, various steps have been made towards establishing a more robust data-based modelling. While, the following text contains mainly peer-reviewed and published work, results under peer review or being prepared for publication are also included to a limited degree.

3.1 Characterization of Asymmetric Laser-induced Plasmas

A crucial factor of analysing any data is the understanding of the underlying processes generating the data. In the context of LIBS, this includes the characterization of the emitting micro-plasma. Much of the theoretical framework used to analyse and model laser-induced plasmas in the context of LIBS is limited to orthogonal ablation geometries using a single ablation laser pulse. However, orthogonal ablation cannot be always ensured. In addition, the possible benefits of using two laser pulses instead of the standard single-pulse LIBS has sparked considerable efforts made towards the understanding of asymmetric laser-induced plasmas as well. The following two peer-reviewed and published works contribute to these efforts.

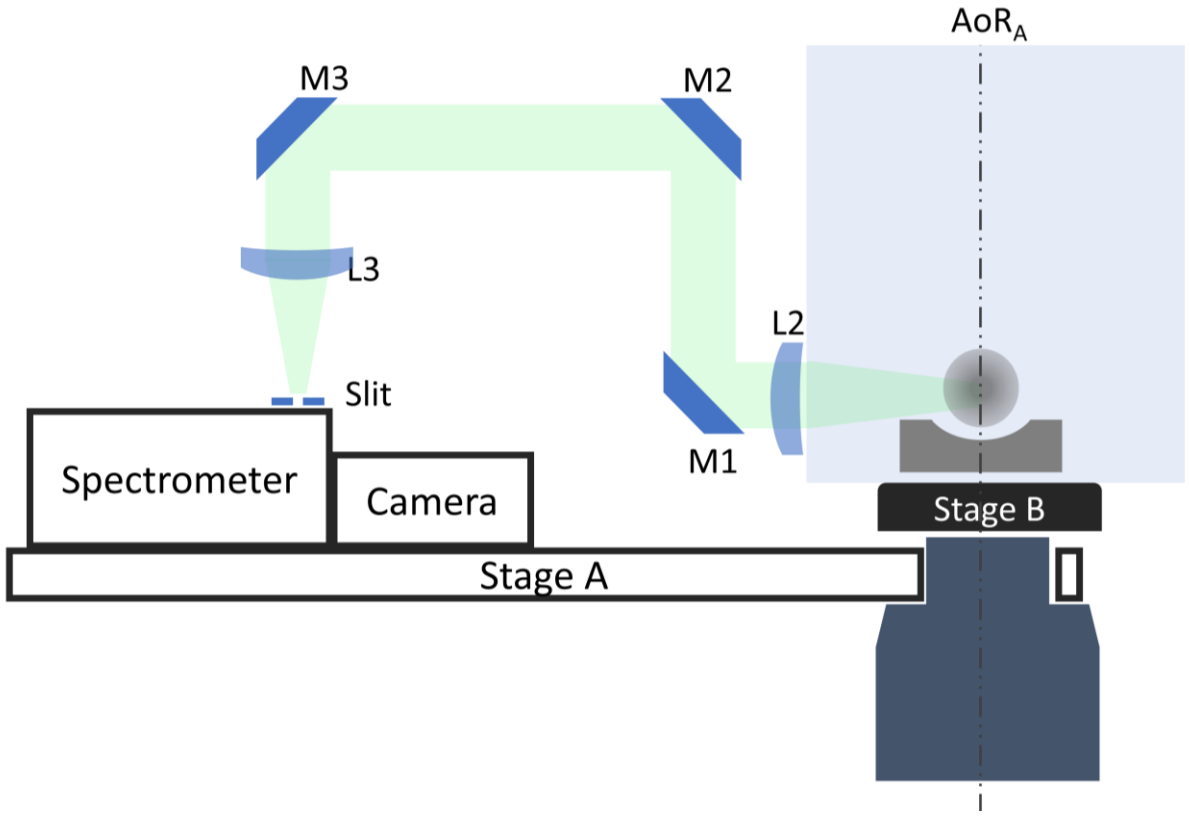

Figure 3-1 Schematic view of the tomographic system. The highlighted portion is modified according to the studied laser-induced plasma. M1—M3: flat mirrors; L2—collection lens with a focal length of f located in a distance of f from the ablation spot; L3—lens imaging the collected light onto the spectrometer's slit; AoR$_A$—axis of rotation of the stage A.

Both works combine the spatially and temporally resolved observation of the LIPs combined with the inverse Radon transform [86] to obtain a detailed three-dimensional distribution of the spectrally resolved, band-filtered, and white-light emissivity; temperature; and electron number density inside the plasma. To carry out the tomographic reconstruction of the plasmas, the world-wide unique instrumentation at the Federal Institute for Materials Research and Testing (Bundesanstalt für Materialforschung und -prüfung; BAM, Berlin, GE) has been used [87]. The instrument is schematically shown in Figure 3-1, with the portion of the system altered for the two works highlighted. The instrument consists of a rotating carousel and a sample manipulator. The carousel (stage A) carries the collection optics, spectrograph, and detector. Meanwhile, the optical system focusing the ablation laser pulse is kept stationary. The sample manipulator ensures that a fresh sample surface is ablated by each laser pulse.

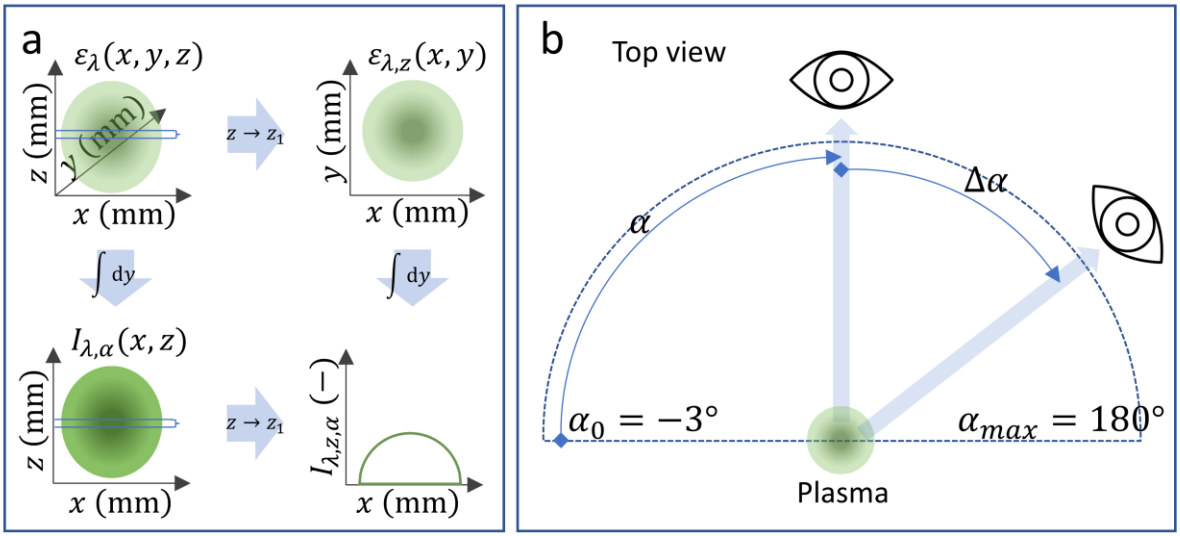

Figure 3-2 a) The emitting plasma volume characterized by a spatial emissivity distribution $\varepsilon_\lambda(x,y,z)$ at the wavelength λ (top left), which is integrated along the line of sight of the detection optics to obtain the recorded image of the plasma $I_{\lambda,\alpha}(x,y)$ at the wavelength λ (bottom left). The complete emissivity distribution can be approximated as a set of discrete thin slices $\varepsilon_{\lambda,z}(x,y)$ (top right) marked by the parallel blue lines in $\varepsilon_\lambda(x,y,z)$. The integral of $\varepsilon_{\lambda,z}(x,y)$ along the line of sight corresponds to a row of the detector recording the plasma's image (bottom right). b) Top view of the tomographic observations: images of the plasma are collected from a set of observational angles α with a step-size of $\Delta\alpha$ in the range of $\{-3, 180\}°$.

The instrument allows the collection of images $I_{\lambda,\alpha}(x,z)$ (Figure 3-2a) of the LIP at various azimuthal observation angles α (Figure 3-2b) with discrete steps $\Delta\alpha$. These images record the intensity of the plasma emission in the form of rows at heights z above the sample surface. The lower index λ refers to the recorded wavelength of the recorded image. These images correspond to the emissivity integrated along the line of sight of the collection optics (the y-axis in Figure 3-2a). Considering the Fourier slice theorem [88] and taking a single row of intensity values $I_{\lambda,z,\alpha}$ collected from $\alpha \in \{0, \dots, 180\}°$ observation angles, the inverse Radon transformation can be applied to recover the emissivity distribution along a slice at the chosen height $\varepsilon_{\lambda,z}(x,y)$. By stacking $\varepsilon_{\lambda,z}(x,y)$, the complete spatial emissivity distribution $\varepsilon_\lambda(x,y,z)$ at a given wavelength λ can be recovered. Alternatively, instead of recording images of the whole plasma at a given wavelength $I_{\lambda,\alpha}(x,z)$, a narrow slit can be applied to select a single slice of the plasma. Subsequently, the collected light can be resolved by a diffraction grating. Thus, by considering the individual wavelength values λ instead of the slice heights z, the spectrally resolved emissivity $\varepsilon_z(x,y,\lambda)$ can be recovered at the axial distance z above the sample surface. Subsequently, the plasma temperature distribution $T(x,y,z)$ (assuming LTE) and

electron number density distribution $N_e(x,y,z)$ can be determined from $\varepsilon_\lambda(x,y,z)$ using the Boltzmann plot technique and the Saha equation, respectively.

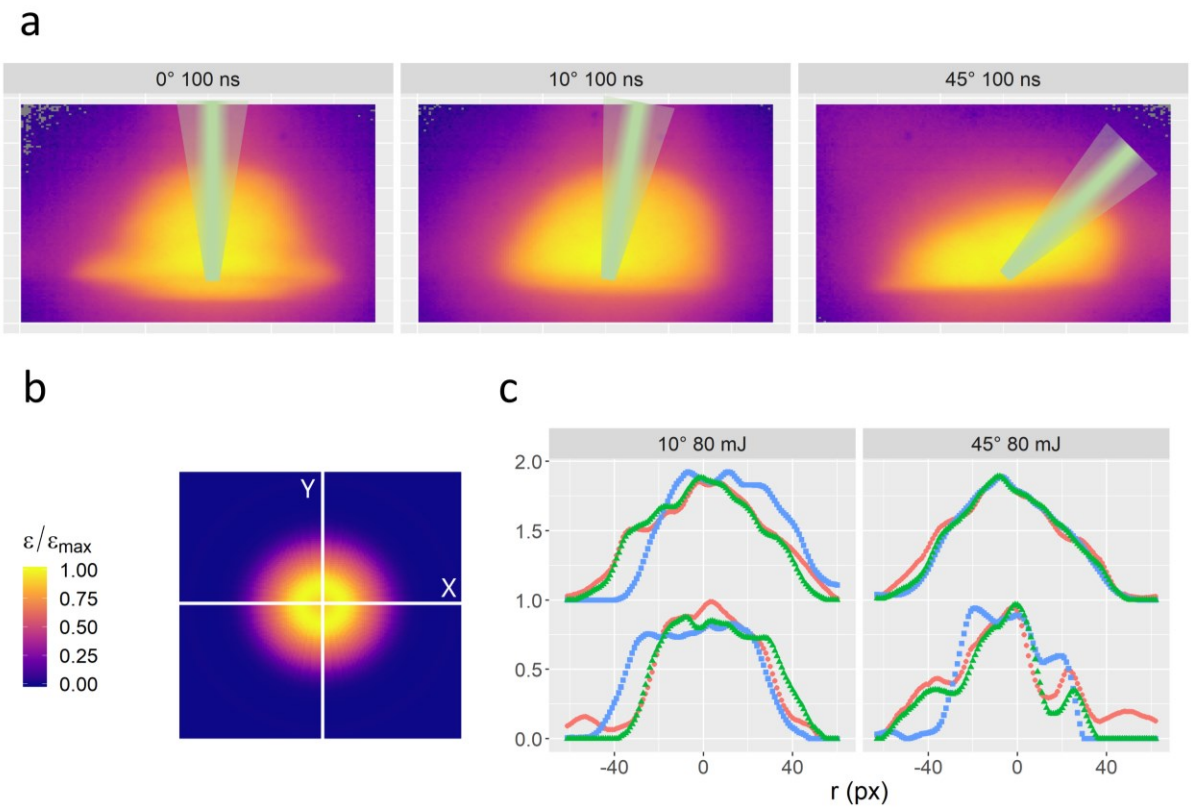

Figure 3-3 Limited summary of [89]. a) According to white-light images, the plasma expands along the ablation pulse. b) Normalized emissivity distribution of the radially symmetric plasma yielded by orthogonal ablation $200\ \mu m$ above the sample surface with two principal directions Y, X shown. c) Normalized Cu I $521\ nm$ emissivity profiles along the principal direction X as observed at three distinct delays (t_d, corresponding to the three distinct colours) at the axial distances z of $200\ \mu m$ (bottom curves) and $500\ \mu m$ (top corves, shifted by 1 for clarity) above the specimen surface. Modified from [89].

In the works [90] and [91], Cu plasmas were characterized considering their overall shape obtained from their white-light emission (i.e., emissivity integrated in the $250—800\ nm$ spectral range) with a fully open slit, spectrally resolved observations at various heights (distances from the specimen's surface) with a narrow slit. The temperature of the plasmas was determined using the Boltzmann plot technique using three atomic emission lines. Subsequently, the electron number density was calculated from Saha's equation using the temperature obtained from the Boltzmann plot technique. Each observation was averaged from 10 ablations (which were recorded individually in single-pulse mode).

3.2 Non-orthogonal Ablation

In many in-situ applications, such as the currently active Mars rovers, the plasma is induced using a non-orthogonal ablation angle. Initial studies of the signal emitted by plasmas induced in conditions relevant to the Curiosity Mars rover showed the strong impact of the ablation angle [92]. Thus, non-orthogonal ablation ought to be studied under measurement conditions relevant to LIBS. In the work titled *"Spatiotemporal spectroscopic characterization of plasmas induced by non-orthogonal laser ablation"* [89], we studied non-orthogonal LIPs with unprecedented spatial and temporal resolution.

The plasmas were induced using an Nd:YAG laser. Three distinct incidence angles of $\{0, 10, 45\}°$ w.r.t. the specimen's surface normal were studied. At each incidence angle, three ablation energies were investigated, namely $\{15, 55, 80\}$ mJ. The ablation pulse was focused using a single lens located at a distance equal to its focal length from the specimen's surface.

The plasmas' white-light emission suggested that the plasmas initially expand along the ablation pulse (Figure 3-3a). However, by using spectrally resolved observations we were able to conclude that the plasma is separated into two parts: a plasma plume consisting of the target material which expands perpendicularly to the specimen's surface and a plasma plume propagating along the laser pulse mainly emitting continuum radiation. Overall, the non-orthogonal ablation angles were observed to have limited impact on the emissivity of atomic species (Figure 3-3), mainly caused by the increased ablation spot size—which in turn lead to the decrease of the laser fluence—resulting from the tilted ablation beam profile. On the contrary, the emissivity of the ionic species exhibited considerable inhomogeneities compared to LIPs produced by orthogonal ablation angles.

Overall, the work succeeded in gaining novel insights into the dynamics of LIPs resulting from non-orthogonal ablation. The obtained results are expected to contribute to a wide range of applications. Nevertheless, the work remains underappreciated by the community. There are several possible reasons behind this. While a robust theoretical understanding of plasma dynamics can help in developing practical applications, it is often more straightforward to build calibration models using data collected in the relevant experimental settings. This is also a likely reason behind the relatively limited efforts made towards the understanding of non-orthogonal ablation in the context of LIBS. Moreover, our results are limited to a relatively simple material, i.e., a certified Cu standard.

3.3 Orthogonal Double-Pulse LIBS

To mitigate some of the limitations of LIBS, the use of multiple laser pulses was early proposed Specifically, the use of two laser pulses is referred to as double-pulse LIBS (DP-LIBS). There are several possible ablation geometries in which DP-LIBS can be performed. The arrangement including a non-orthogonal laser pulse produces radially asymmetric LIPs which remain largely unexplored. Thus, in the work titled *"Tomography of double-pulse laser-induced plasmas in the orthogonal geometry"* [91] we carried out the detailed comparison between the two ablation geometries available for orthogonal DP-LIBS in terms of temporally resolved three-dimensional distribution of the emissivity of atomic and ionic species, plasma temperature, and electron number density.

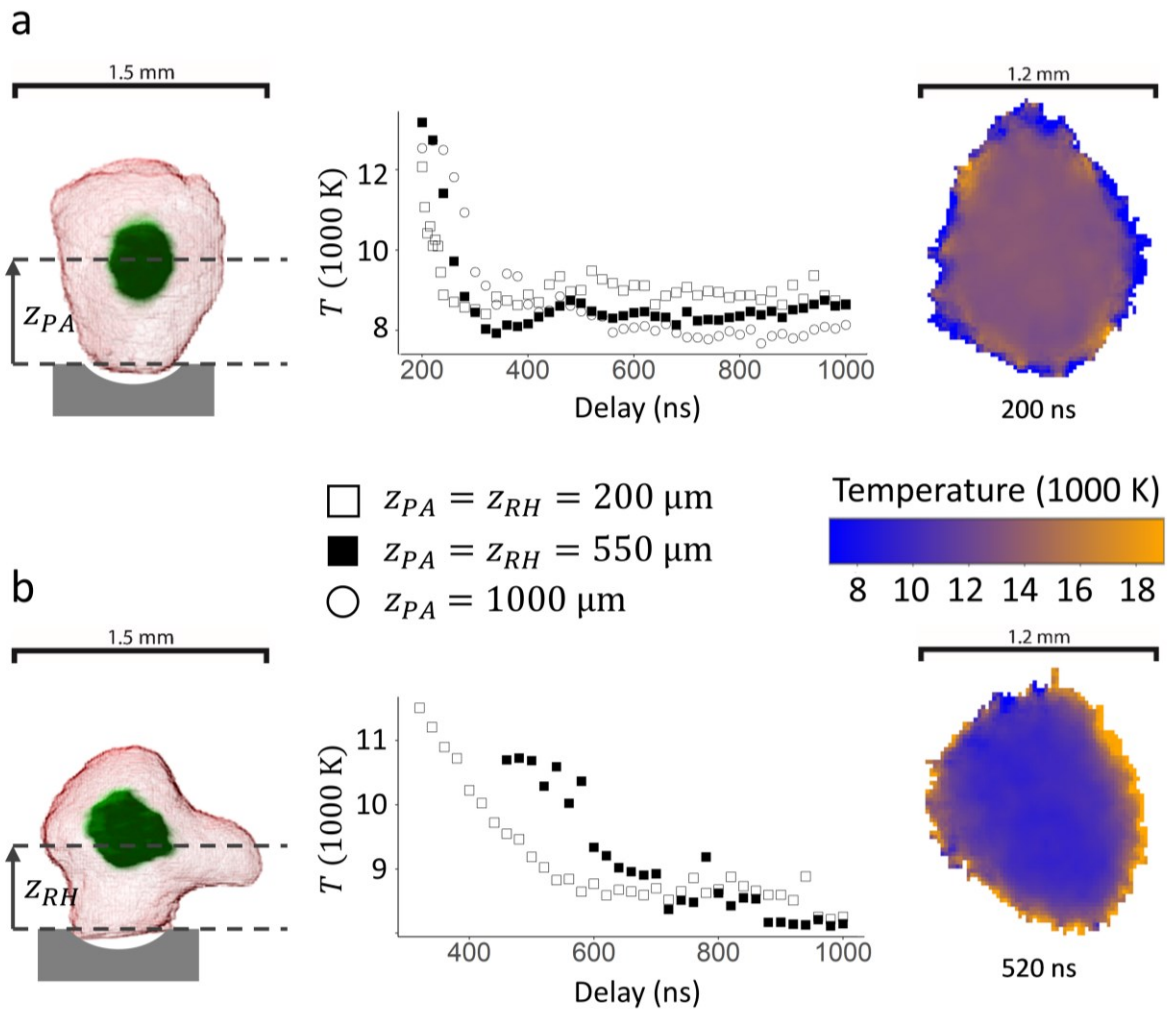

Figure 3-4 Limited summary of [91]. a) White-light spatial emissivity distribution of the plasma yielded by orthogonal pre-ablation (left); the corresponding temporal evolution of the temperature at distinct z_{PA} axials distances from the specimen's surface (middle); and the corresponding spatial temperature distribution at $z_{PA} = 1000$ µm (right). b) White-light spatial emissivity distribution of the plasma yielded by orthogonal re-heating (left); the corresponding temporal evolution of the temperature at distinct z_{RH} axials distances from the specimen's surface (middle); and the corresponding spatial temperature distribution at $z_{RH} = 550$ µm (right). Modified from [91].

The plasma expansion velocity was determined by reconstructing the white-light emissivity of the plasmas at various time-steps. Nevertheless, the white-light emissivity also included plasma regions which did not contain specimen material, e.g., the air spark generated by the pre-ablation pulse. Thus, an optical bandpass filter (515—525 nm spectral range) was used to acquire images of the full plasma in the spectral range dominated by the emission lines of Cu. Consequently, we were able to distinguish between the expansion velocity of the specimen material and that of the whole plasma. In addition, the spatial homogeneity and temporal stability of the plasmas' electron number density and temperature were explored at various axial distances from the specimen's surface.

We observed that the pre-ablation geometry yields a temporally more stable plasma. According to the current understanding of orthogonal pre-ablation, this is the result of the smaller resistance of the ambient atmosphere resulting from the air spark created above the sample surface. On the contrary, in limited temporal windows, the re-heating geometry was observed to yield higher spatial homogeneity compared to both the pre-ablation DP geometry and the common SP ablation. The

overall impact of the work remains to be seen. Nevertheless, our results have already attracted some attention, with a few works building on them.

3.4 Reconsidering Spectra Acquisition Strategies

One of the most common assumptions made by statistical models is that the data is drawn from a normal distribution. However, it has been repeatedly reported that LIBS data do not follow a normal distribution [93]. In addition, it has been shown that the data acquisition strategy (the number of spectra collected from a single spot, the number of plasmas whose emission is integrated to obtain a single spectrum, etc.) can influence the distribution of the data. Thus, in the work titled *"On the application of bootstrapping to laser-induced breakdown spectroscopy data"* [94] we explored various data acquisition strategies. Namely, we compared raw single-pulse measurements with measurement accumulated from a varying number of laser shots and single-pulse measurements processed using bootstrapping. Bootstrapping refers to the repeated resampling of the dataset with replacement and subsequently averaging the sampled data points to obtain new data points.

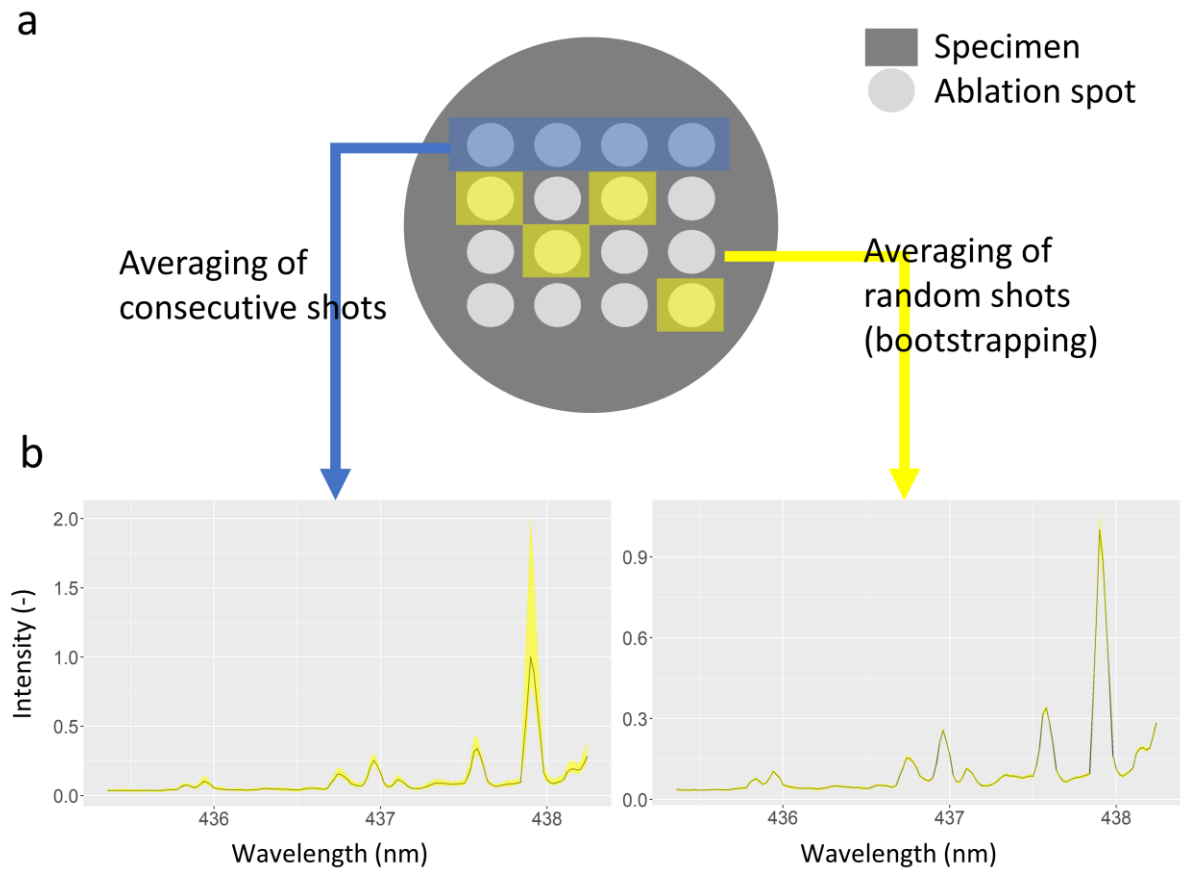

Figure 3-5 Limited summary of [95]. a) The specimen was sampled multiple times and the plasma's emission was collected considering distinct strategies. b) Mean emission spectrum (in limited spectral range) of an aluminium specimen obtained by averaging consecutive single-shot spectra (left) and averaging randomly selected single-shot spectra (right). The Modified from [95].

Three different sample matrices were studied, namely steel, bronze, and aluminium. The specimen surfaces were treated in the same way to eliminate the influence of the surface topography. Then, signal from 10000 shots were collected following different strategies: 1) single pulse; 2) accumulation of echellograms prior to their conversion to broad-band spectra; 3) numerically accumulated consecutive single-pulse spectra; and 4) bootstrapped single-pulse spectra. Consequently, the spectra

were evaluated in terms of relative standard deviation, signal-to-background and signal-to-noise ratios, and probability distribution of the emission line intensities. During the analysis, major and minor constituents of the samples were considered.

According to our results, it is more beneficial to collect single-shot spectra and consequently process them using bootstrapping than to accumulate signal from consecutive ablations (Figure 3-5). Bootstrapping is especially beneficial for obtaining data that follow the normal distribution. Many of the benefits of bootstrapping result from the central limit theorem which states that repeatedly (statistically) sampling an unknown probability distribution and averaging the sampled data points yields a normal distribution. Moreover, increasing the number of samplings decreases the standard deviation of the final Gaussian distribution. However, the mean of the real and the bootstrapped distributions remains identical. Thus, while bootstrapping can be used to homogenize the distribution of LIBS datasets, it should not be used to expand measured datasets with the aim of providing novel data points. The latter process is referred to as data augmentation and is commonly used to build more robust machine learning models with minimal cost of collecting new data. Nevertheless, bootstrapping is not an appropriate technique for data augmentation.

3.5 Optimizing the Spectral Pre-processing

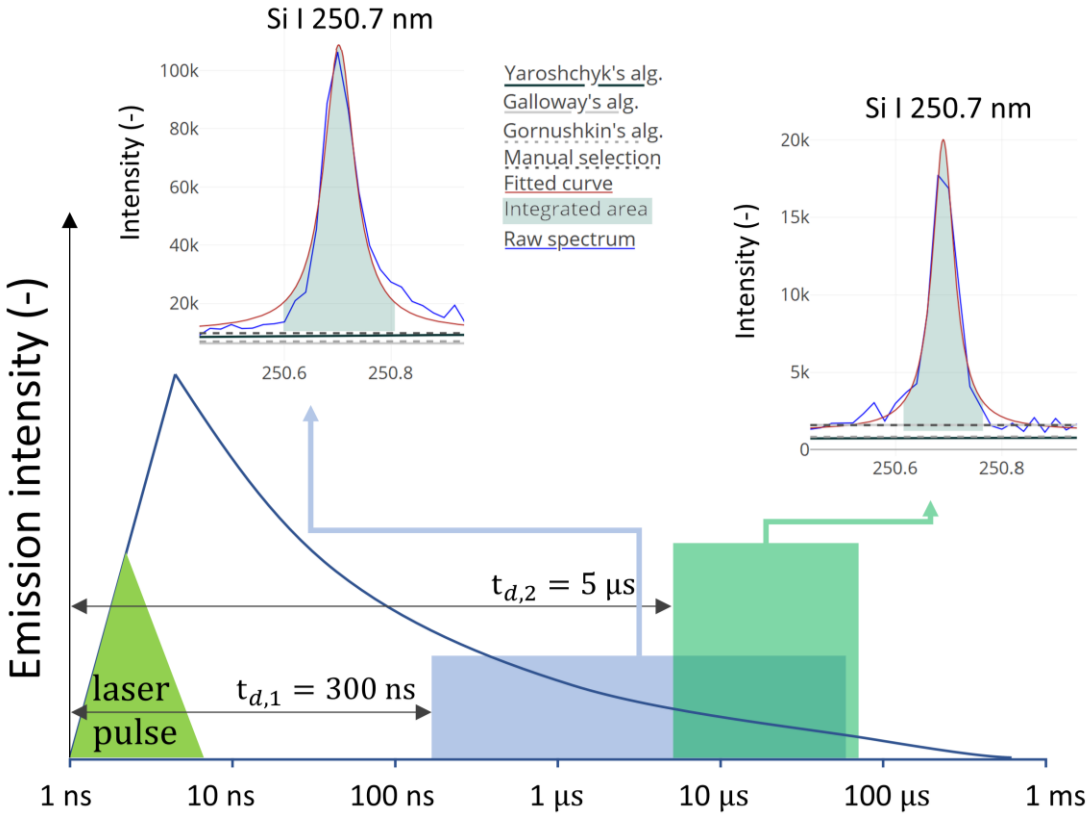

Figure 3-6 Spectra were collected at distinct delays (t_d) with the same gate width ($t_w = 50$ µs). Subsequently, four different baseline correction algorithms were applied. Yaroschchyk's algorithm utilizes a smoothed moving average filter; Galloway's algorithm relies on wavelet denoising; Gornushkin's algorithm applied polynomial fitting of the local minima; and manual baseline selection refers to the manual selection of the baseline's magnitude. After baseline correction, the emission line's intensity was obtained as the area under the fitted Lorentz profile. Modified from [96].

The background signal present in LIBS spectra is often removed numerically in a pre-processing step called baseline correction. Owing to the variety of techniques available for carrying out baseline correction, the performance of these techniques should be evaluated considering the application at

hand. Consequently, the work titled *"Influence of baseline subtraction on laser-induced breakdown spectroscopic data"* [96] carried out such a comparison with regards to the baseline correction's impact on the limit of detection of various minor constituents of a certified aluminium standard.

Spectra were collected at various delays. Consequently, the spectra exhibited various levels of background signal, i.e., baseline. In addition, the spectra also contained various levels of analyte signal. Subsequently, four distinct baseline correction algorithms were applied to the spectra and the limit of detection was determined for every analyte contained in the sample. The limit of detection (LOD) obtained from the treated and raw spectra were compared. The baseline correction approaches included the manual selection of the baseline's level, the polynomial fitting of the spectrum's local minima, wavelet denoising, and a smoothed moving minimum operation.

The obtained results (Figure 3-6) presented an alternative to the common paradigm. In general, the spectral baseline is suppressed by adjusting the measurement parameters such as the delay of the acquisition. Instead, according to the work, the baseline can be numerically treated in spectra recorded with short gating settings (at the early stage of the plasma, when it is characterized by a strong continuum background radiation). This observation has been since used to improve the analytical performance of LIBS considering the detection of minor elements. Nevertheless, an overall best-performing baseline correction algorithm could not be identified. Moreover, similar studies involving several distinct matrices would be highly beneficial and might provide additional insights.

3.6 Taking Advantage of the Sparsity of LIBS Data

One of the most popular statistical tools used in the LIBS literature is PCA [51]. The embedding yielded by PCA often provides a reasonable interclass separation while maintaining intraclass cohesion, thus, aiding both quantitative and qualitative analyses. Moreover, PCA is easily interpretable. Nevertheless, PCA scales unfavourably with the ever-increasing size of LIBS datasets. Thus, in the work titled *"Addressing the sparsity of laser-induced breakdown spectroscopy data with randomized sparse principal component analysis"* [97] two extensions of PCA were studied in the context of LIBS. Namely, sparse PCA (SPCA) was explored to explicitly address the sparsity of LIBS data for providing improved interpretability of the PCA loading spectra. In addition, randomized SPCA (RSPCA) was explored as a possible acceleration of the traditional PCA.

Sparse PCA includes the L1 norm in its optimization problem, which results in a sparse solution. In comparison to the loading spectra yielded by the traditional PCA, the L1 norm greatly reduces the number of non-zero covariate weights (Figure 3-7a). Thus, fewer emission lines are found to be relevant, leading to a less demanding emission line identification process. Meanwhile, comparable clustering is obtained as with PCA (Figure 3-7b). Apart from the L1 norm, the inclusion of the L2 norm reduces the magnitude of the weights and forces the weights of covariates exhibiting high covariance to be similar.

The magnitude of the constants determining the prevalence of the two regularization terms in the optimization problem can lead to a slightly different embedding. Thus, the work also explores the possibility of combining several SPCA models trained with different L1 and L2 regularization magnitudes to enhance the clustering performance on the embedded data points. To find optimal conditions, genetic algorithms were used. Nevertheless, the approach proved to be a dead end; the high computational demands of the genetic algorithm did not justify the minute gains in the clustering quality compared to a simple SPCA model.

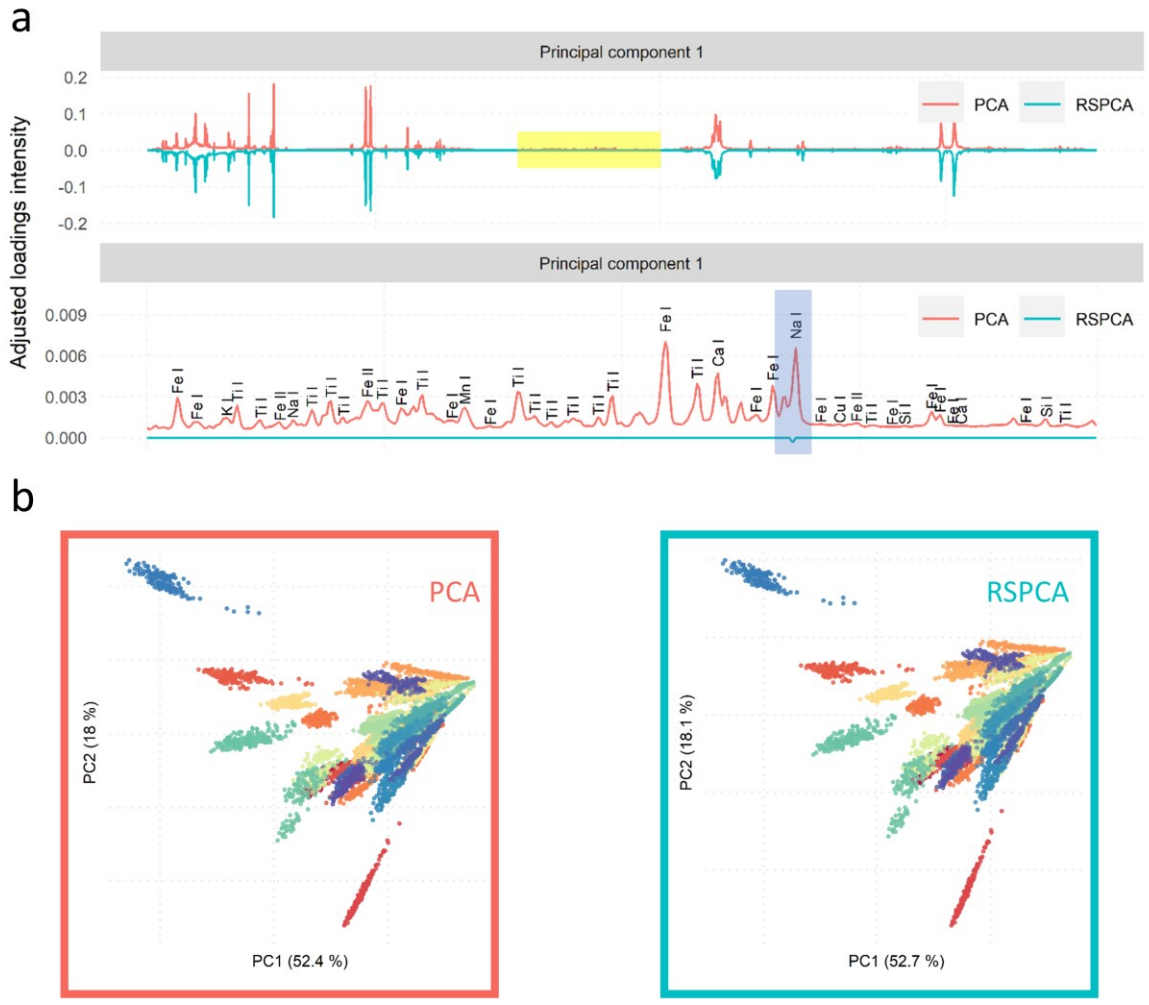

Figure 3-7 Limited summary of [97]. A dataset was projected into a low-dimensional feature space using the loadings (covariate weights) shown in a) to obtain the embeddings shown in b). PCA—principal component analysis; RSPCA—randomized sparse PCA. The emission line highlighted with blue in a) shows the only non-zero covariate weights yielded by RSPCA in the spectral region highlighted by yellow. The percentage points denote the percentage of variance in the dataset explained by the corresponding principal component (PC). Modified from [97]

The second modification of the initial PCA optimization problem made by RSPCA is the application of randomized linear algebra. Namely, since LIBS data matrices are sparse, the QR decomposition of a matrix with a reduced size can carry enough information of the initial dataset[98]. Consequently, the matrix decomposition—a fundamental part of the PCA algorithm—is significantly simplified. This leads to the 20-fold acceleration of PCA.

Overall, the considerable acceleration of PCA alone is a valuable observation. In addition, our work provided a range of well-performing hyperparameters for applying RSPCA to LIBS data. Moreover, the benefits of the sparse loading spectra were demonstrated. Lastly, the work compared the loading spectra yielded by SPCA to those yielded by sparse PLS regression (SPLSR). The incorporation of the same L1 regularization term into the PLS optimization problem has been used to study the impact of matrix effects on various occasions. Nevertheless, our comparison showed that owing to the unsupervised nature of the SPCA algorithm, the yielded loading spectra do not carry the same information as the loadings obtained from the SPLSR, which is built in a supervised manner.

3.7 Beyond Linear Dimensionality Reduction

Despite dimensionality reduction being ubiquitous in LIBS analysis, the developments made in the past few years regarding DR have been minimal in the LIBS literature. Thus, most works rely solely on PCA. This poses severe limitations to the development of the LIBS methodology. A crucial factor defining the potential success or shortfall of machine learning models is the representation of the data. Considering the various characteristics of LIBS spectra, e.g., the non-linear signal response and sparsity, PCA is expected to be outperformed by more advanced DR techniques. However, while most advanced DR techniques initially outperform PCA on synthetic toy datasets, they often exhibit subpar performance on real datasets [50]. This prompts a detailed investigation and comparison of the most popular DR techniques. Consequently, the work titled "*Comprehensive evaluation of the dimensionality reduction of spectroscopic data*" (to be submitted for publication) evaluates the quality of the low-dimensional embedding yielded by several non-linear DR approaches. Adequate clustering in a low-dimensional space is indicative of both the performance of qualitative and quantitative analysis.

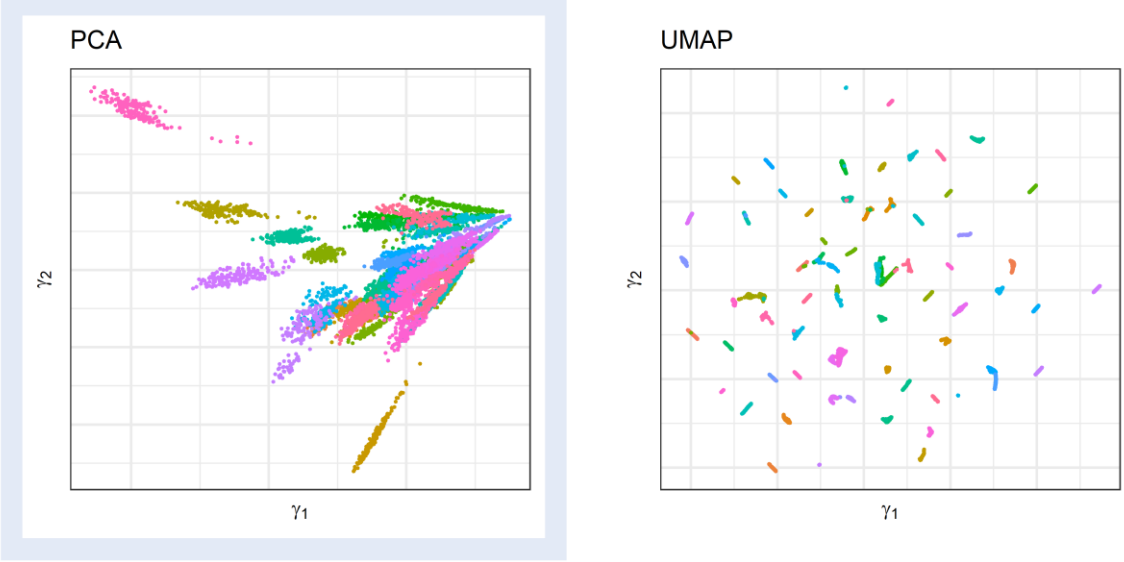

Figure 3-8 Selected examples of the non-linear low-dimensional embedding of the ChemCam calibration dataset (over 11000 spectra from 67 specimens) [99]. PCA—principal component analysis (linear embedding for reference); UMAP—uniform manifold approximation and projection [100].

A comparison of PCA with a non-linear embedding is shown in Figure 3-8. The preliminary results show that PCA—despite limited to linear transformations—performs comparably to most non-linear techniques. On the contrary, a few recent DR techniques (such as UMAP [100]) clearly outperform PCA, although, at the cost of the complete loss of interpretability. The metrics used to compare the DR techniques enumerate the separability of the clusters and the cohesiveness of the data points inside the individual clusters.

3.8 Choosing the Right Classification Model

The only currently available methodology for identifying the best-performing model is to treat the model itself as a hyperparameter that must be optimized. Consequently, several models are fine-tuned using CV and the final model is selected according to their test performance. Choosing what pre-processing to carry out prior to the analysis poses similar challenges. Thus, it is not surprising that the development of a novel analysis technique claiming SOTA results is not uncommon in the current

LIBS literature. However, most of these works lack a convincing proof of such claims. Namely, the models and the preceding spectral pre-processing are rarely evaluated on publicly available datasets. Consequently, an extensive dataset has been prepared and subjected to an international classification challenge to objectively compare various data analysis pipelines. After the contest's conclusion, the dataset was made publicly available.

The dataset and its preparation are detailed in the data descriptor titled *"Benchmark classification dataset for laser-induced breakdown spectroscopy"* [101]. The aim of the dataset is to enable LIBS users to compare and evaluate newly developed classification methodologies and to trigger numerous recent breakthroughs similarly to the field of machine learning, where the compilation of a range benchmark datasets sparked advances in many applications. Thus, the dataset was designed to make the classification task adequately challenging. Consequently, we deviated from the usual in-sample classification task and opted for out-of-sample classification. In-sample classification refers to the random division of an available dataset into training, validation, and testing subsets. This is frequently misinterpreted in the LIBS literature. Namely, LIBS datasets are erroneously randomly divided in a manner that includes spectra from the same specimens in both the training and testing subsets. Since LIBS data of homogeneous samples commonly follow a relatively simple probability distribution, this leads to a relatively simple classification task. Consequently, the reported test accuracies are unrealistically high, and the models would probably fail on spectra collected from new specimens made of the same material.

Meanwhile, in practical applications (which LIBS is generally designed for), a degree of generalizability of the classification model is required. As an example, in the case of a LIBS instrument installed in a metal-sorting facility, spectra of dozens of specimens might be available for training a classification model. However, the trained model is then expected to correctly classify thousands of new specimens, often contaminated, and exhibiting a mildly altered composition. Thus, the appropriate way of constructing datasets is to collect the training and testing data from distinct specimens. In turn, the testing performance of the classification models is evaluated out-of-sample, that is, on a differently sampled dataset than the training dataset. This was the motivation behind the constructed dataset. The dataset consists of 12 classes and 138 specimens (unevenly distributed among the 12 classes). Every specimen has a unique chemical composition.

The dataset has been the subject of an international classification contest. The contest consisted of two parts: 1) the contestants could continuously test their model's performance on a test dataset, which remained unchanged; 2) in a second round, the classification of a new set of spectra from the same set of testing specimens was attempted. While the participation in the contest's first round has been reasonable, some of the best teams did not attempt the second round. The results of the contest and the gained insights are summarized in the accompanying publication titled *"Classification of challenging Laser-Induced Breakdown Spectroscopy soil sample data - EMSLIBS contest"* [53].

The methodologies applied during the contest ranged from relatively simple linear models to SOTA non-linear ANNs. However, every approach composed of roughly the same steps: finding and appropriate representation of the data followed by the fine-tuning of a selected classification model. The winning team (of the first round[2]) used a linear multivariate transformation followed by a linear classification model. However, they complemented the simple models with extensive spectroscopic expertise and the laborious pre-processing of the dataset. On the contrary, most other teams

[2] The team did not participate in the second round.

employed highly automatized data-processing workflows for selecting meaningful features. In addition, several teams applied SOTA dimensionality reduction techniques (e.g., UMAP [100]).

Several conclusions could be drawn from the classification contest. 1) Spectroscopic expertise combined with the careful analysis of LIBS data can still outperform machine learning approaches. 2) Blindly following machine learning techniques can lead to overfitting and poor generalization capabilities, which was shown by the significant drop in classification accuracy in the second round of the contest by multiple teams. 3) Splitting the training data for CV in a manner that imitates the out-of-sample nature of the classification task greatly improved the generalizability of ML models. Overall, the compilation and comparison of vastly distinct approaches to the analysis of LIBS spectra has been welcomed by the community.

3.9 Understanding Black-Box Multivariate Models

One of the major findings of the classification contest has been the tendency of LIBS specialists to overtrain multivariate models. The severity of this issue increases with the capabilities of the used technique. Coincidentally, most ML models exhibiting SOTA performance are opaque or black-box models. In addition to the risk of overtraining, the lack of understanding of black-box models also limits their performance. Hence, modern data-based modelling approaches are unlikely to solve the limitations of LIBS unless they can be properly understood. Thus, the interpretability of common black-box models is explored in a series of works currently under peer review and under preparation for publication. Lastly, the understanding of black-box models applied o LIBS data is also urged by LIBS gaining prominence in high-stakes applications, such as being used as a complementary technique for the analysis of cancerous tissues.

3.9.1 Interpretation of Support Vector Classifiers

One of the most frequently used classification models in the LIBS literature are SVCs. However, SVCs do not provide loading spectra like PLS, which would be indicative of the features importance. Thus, SVCs are black box models. Consequently, the identification of the features considered important by SVCs has been attempted. However, the approaches present in the LIBS literature do not directly investigate the importance of features or covariates for the model. Rather, they generally perform the iterative evaluation of SVC models with a subset of the available covariates. The subset of covariates yielding the best classification results is than regarded as the most important ones. However, since the models are retrained (and optimized) on each subset, these results are not indicative of the importance of a feature/covariate for a specific model. Rather, it provides insights into which covariates carry the most information about the difference between samples.

Consequently, in the work titled "*Interpreting support vector machines applied in laser-induced breakdown spectroscopy*" currently under peer-review, several approaches to interpreting SVCs were explored. Namely, the work compared four methodologies of determining which covariates are considered by SVCs for the classification. The fundamental idea of the four approaches is to probe the trained classification model by a modified training dataset. That is, the model is used to classify the modified training dataset and the predictions yielded for the unperturbed and modified datasets are compared. The four approaches differ in the methodology used for obtaining the probe dataset and the metric used to evaluate the resulting change in the model's prediction.

The findings of the work suggest that a set of feature importance metrics should be used for the complete analysis of SVCs as each provide a distinct set of important features. Similar feature importances were observed for linear-kernel and non-linear kernel SVCs, suggesting that LIBS data

commonly do not benefit from kernel transformations into a higher dimensional space for classification. In other words, the representation of LIBS data in their initial feature space already allows linear separability. Most importantly, the work shows that SVCs are able to identify whole spectroscopic features even if the classification of complete spectra are considered. Considering that only limited correlation was observed between the features considered important by the SVC models and covariate weighting done by PCA, the interpretation of SVCs offers a novel approach to finding informative features in large LIBS datasets.

3.9.2 Interpretation of Convolutional Neural Networks

The breakthroughs achieved by the introduction of ANNs proliferated into most scientific fields. Nevertheless, these models are notoriously difficult to interpret. This led to the whole research field aimed at the development of methods providing insight into the working principles of ANNs. In the work titled "*Towards interpretable convolutional neural networks applied to spectroscopic data*" currently being prepared for publication, several of these methods are adjusted to LIBS data to investigate how the analysis of spectroscopic data can benefit from CNNs. Such insights can help develop more robust models, cut time and monetary costs of fine-tuning, and provide trust.

Namely, a CNN model is trained to classify the extended ChemCam calibration dataset (over 400 specimens). Subsequently, the shape of the filters of the trained model are investigated. In addition, the convolutional transformation of the data is investigated, i.e., the internal representation created by the convolutional filters is visualized. Moreover, the ability of CNNs to identify spectroscopic features rather than individual covariates is explored. Lastly, a set of ideal spectra are obtained *via* gradient ascent optimization which maximize the activation of the output neurons, i.e., spectra which correspond to the perfect classification to the corresponding class.

Overall, our results suggest that despite the lack of spatial invariance of LIBS data, CNNs (which were designed to take advantage of patterns exhibiting spatial invariance) provide multiple benefits. Namely, the internal representation of the trained CNN exhibits better interclass separation. In addition, CNNs—similarly to SVCs—can identify whole spectroscopic features. The work's findings will allow a more directed CNN architecture choice for LIBS applications and the verification of the robust decision making of trained models prior to their application in practical settings.

Summary and Conclusions

Laser-induced breakdown spectra are generated *via* relatively simple and well-understood processes. The optical emission of LIPs can be well-described using a few number densities (those of electrons, atoms, ions), and under optimal plasma condition—i.e., LTE—using a single temperature. However, LIPs evolve in time. Moreover, the LIPs' temporal evolution is non-linear and can be accurately modelled only by solving the Navier—Stokes equations. Consequently, minor changes in the initial conditions of the LIPs' formation considerably affect the LIPs' optical emission integrated over the time scales commonly considered for LIBS. Consequently, various matrix effects result in considerable changes in LIBS spectra.

More importantly, changing experimental conditions—including the instrumental variability stemming from LIBS's flexibility—can yield incomparable datasets from even the same set of specimens. Some of the major impacts result from the convoluted dynamics of laser ablation; the laser pulse's wavelength and duration have fundamental effect on the ablation mechanisms. Moreover, due to the spatial inhomogeneities of LIPs, the collection optics can also alter the obtained spectra.

The experimental and theoretical studies available of the above-mentioned effects are abundant. Nevertheless, a systematic comparison of the various experimental factors is missing; Often, studies using different experimental setups are compared to draw conclusions, disregarding the considerable impact of several factors. Thus, further studies are necessary before the general transfer of LIBS spectral libraries can be developed. Namely, conscious efforts must be made towards the parametrization of the impact of the individual instrumental parameters and their convoluted effect, which are currently expressed mainly empirically. A potentially successful parametrization could be obtained using the data-based modelling approaches developed in the past decade by the field of machine learning (ML). The existing effort towards the use of ML to achieving LIBS spectral library transfer are the subject of Part II of this thesis.

The transfer of LIBS spectral libraries across distinct experimental conditions and LIBS systems can be regarded as a specific case of transfer learning, a subfield of machine learning research. Transfer learning aims at utilizing data obtained from different sources to enhance the performance of (machine learning) calibration models applied to a specific dataset. In fact, transfer learning has already been applied to adapt to changing sample properties. In the context of near-infrared spectroscopy, transfer learning has been applied also to adapt to changing measurement parameters. While near-infrared spectra do not exhibit the same variance resulting from changing measurement systems as LIBS, the successful application of transfer learning in near-infrared spectroscopy suggests the viability of the transfer learning approaches. However, owing to the unique nature of LIBS data, the currently available transfer learning methodologies are not directly applicable to LIBS data.

Among the two general directions of transfer learning (data-based and model-based), data-based approaches are currently better developed in LIBS. Namely, the successful applications of LIBS all implemented data-based transfer learning. In addition, the rich literature available on LIBS spectral normalization can be also regarded as efforts made towards the standardization of LIBS spectra, although limited to less prominent signal fluctuations than those resulting from the change of measurement conditions.

The potential of model-based transfer learning should not be neglected either. The tools developed by in the ongoing revolution of data processing whose aim is to parameterize complex functions

without explicit programming (i.e., machine learning models) are starting to get traction in the LIBS literature. Currently, in the context of LIBS, the application of machine learning methods is limited to regression and classification analyses. In addition, owing to the black box nature of these models, their objective comparison is lacking. However, the ability to parametrize any function provided with enough data is a promising approach to adapt LIBS spectral libraries to the changing experimental conditions. Nevertheless, currently no efforts have been proposed to measure the required dataset which would encompass a wide range of specimen matrices measured under a variety of experimental conditions.

Owing to the non-linear signal response of LIBS, these models are often complicated and require a vast set of specimens to construct. Hence, the construction of robust and well-performing calibration models for LIBS generally incurs considerable costs. Meanwhile, owing to the flexibility and variability of LIBS systems, combined with the sensitivity of LIBS data to the experimental conditions, the universality of LIBS calibration models is severely limited.

The methodologies developed by the field of transfer learning offer a viable solution. Namely, data-based modelling approaches—i.e., machine learning—could be adjusted to the changing measurement conditions and instrumentation using a limited set of calibration specimens. This approach would make the use of existing calibration models for new analyses viable at the fraction of the cost of constructing new models from scratch.

The present thesis reviewed the state of such transfer learning approaches in the context of spectroscopic analysis and identified the possible path towards a more general framework for transfer learning for LIBS. Several challenges have been identified, multiple of which have been at least partially addressed, yielding six published peer-reviewed first-author papers, one additional first-author paper currently under review, and two papers under preparation for publication at the writing of these words. These works contributed towards a more complete understanding of asymmetric laser-induced plasmas, which are commonly encountered by both *in-situ* and laboratory LIBS analyses. Moreover, the work also provided novel insights concerning the spectral acquisition strategies used in LIBS. In addition, the work contributed to the design and preparation of an extensive benchmark dataset for LIBS classification models that was the subject of an international classification contest. The latter helped identifying the current SOTA LIBS classification methodologies. Considering the findings of this work, the several studies of the dimensionality reduction of LIBS data commenced. One of the studies exploited the sparsity of LIBS spectra to accelerate PCA applied to LIBS datasets. Another related work explored non-linear alternatives to PCA. Lastly, as the black-box nature of most machine learning models was identified as a significant limitation of their applicability for transfer learning, several works aimed at the interpretation of common machine learning models are currently being peer-reviewed or prepared for publication.

Overall, the transferability of LIBS calibration models is still in its infancy. Nevertheless, the thesis contributed towards paving the path towards successful transfer of machine learning models across distinct LIBS systems and measurement condition.

Author's own publications

[1] **E. Képeš**, J. Vrábel, P. Pořízka, J. Kaiser, Addressing the sparsity of laser-induced breakdown spectroscopy data with randomized sparse principal component analysis, *J. Anal. At. Spectrom.* (2021) Advance Article.

Author's contribution: 80%. Number of citations: 0 (19.10.2021).

[2] **E. Képeš**, I. Gornushkin, P. Pořízka, J. Kaiser, Spatiotemporal spectroscopic characterization of plasmas induced by non-orthogonal laser ablation, *Analyst*. **146** (2021) 920–929.

Author's contribution: 80%. Number of citations: 1 (19.10.2021).

[3] **E. Képeš**, I. Gornushkin, P. Pořízka, J. Kaiser, Tomography of double-pulse laser-induced plasmas in the orthogonal geometry, *Anal. Chim. Acta.* **1135** (2020) 1–11.

Author's contribution: 80%. Number of citations: 3 (19.10.2021).

[4] J. Vrábel, **E. Képeš**, L. Duponchel, V. Motto-Ros, C. Fabre, S. Connemann, F. Schreckenberg, P. Prasse, D. Riebe, R. Junjuri, M.K. Gundawar, X. Tan, P. Pořízka, J. Kaiser, Classification of challenging Laser-Induced Breakdown Spectroscopy soil sample data - EMSLIBS contest, *Spectrochim. Acta - Part B At. Spectrosc*. **169** (2020) 105872.

Author's contribution: 5%. Number of citations: 13 (19.10.2021).

[5] **E. Képeš**, J. Vrábel, S. Střítežská, P. Pořízka, J. Kaiser, Benchmark classification dataset for laser-induced breakdown spectroscopy, *Sci. Data*. **7** (2020) 53.

Author's contribution: 35%. Number of citations: 7 (19.10.2021).

[6] **E. Képeš**, P. Pořízka, J. Kaiser, On the application of bootstrapping to laser-induced breakdown spectroscopy data, *J. Anal. At. Spectrom*. **34** (2019) 2411-2419.

Author's contribution: 70%. Number of citations: 3 (19.10.2021).

[7] P. Pořízka, J. Klus, **E. Képeš**, D. Prochazka, D.W. Hahn, J. Kaiser, On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review, *Spectrochim. Acta - Part B At. Spectrosc*. **148** (2018) 65–82.

Author's contribution: 15%. Number of citations: 88 (19.10.2021).

[8] **E. Képeš**, P. Pořízka, J. Klus, P. Modlitbová, J. Kaiser, Influence of baseline subtraction on laser-induced breakdown spectroscopic data, *J. Anal. At. Spectrom*. **33** (2018) 2107–2115.Author's contribution: 52%. Number of citations: 8 (19.10.2021).

[9] P. Pořízka, J. Klus, D. Prochazka, **E. Képeš**, A. Hrdlička, J. Novotný, K. Novotný, J. Kaiser, Laser-Induced Breakdown Spectroscopy coupled with chemometrics for the analysis of steel: The issue of spectral outliers filtering, Spectrochim. *Acta Part B At. Spectrosc*. **123** (2016) 114–120.

Author's contribution: 5%. Number of citations: 30 (19.10.2021).

References

[1] Musazzi, S. and U. Perini. *Laser-Induced Breakdown Spectroscopy*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. Springer Series in Optical Sciences. ISBN 978-3-642-45084-6. Available at: doi:10.1007/978-3-642-45085-3

[2] Winefordner, JD., IB. Gornushkin, T. Correll, E. Gibb, BW. Smith and N. Omenetto. Comparing several atomic spectrometric methods to the super stars: special emphasis on laser induced breakdown spectrometry, LIBS, a future super star. *Journal of Analytical Atomic Spectrometry*. 2004, **19**(9), 1061. ISSN 0267-9477. Available at: doi:10.1039/b400355c

[3] Galbács, G. A critical review of recent progress in analytical laser-induced breakdown spectroscopy. *Analytical and Bioanalytical Chemistry*. 2015, **407**(25), 7537–7562. ISSN 1618-2642. Available at: doi:10.1007/s00216-015-8855-3

[4] Yu, X., Y. Li, X. Gu, J. Bao, H. Yang and L. Sun. Laser-induced breakdown spectroscopy application in environmental monitoring of water quality: a review. *Environmental Monitoring and Assessment*. 2014, **186**(12), 8969–8980. ISSN 0167-6369. Available at: doi:10.1007/s10661-014-4058-1

[5] Sturm, V. and R. Noll. Laser-induced breakdown spectroscopy of gas mixtures of air, CO_2, N_2, and C_3H_8 for simultaneous C, H, O, and N measurement. *Applied Optics*. 2003, **42**(30), 6221. ISSN 0003-6935. Available at: doi:10.1364/AO.42.006221

[6] Margetic, V., M. Bolshov, A. Stockhaus, K. Niemax and R. Hergenröder. Depth profiling of multi-layer samples using femtosecond laser ablation. *J. Anal. At. Spectrom.* 2001, **16**(6), 616–621. ISSN 0267-9477. Available at: doi:10.1039/B100016K

[7] García, CC., M. Corral, JM. Vadillo and JJ. Laserna. Angle-Resolved Laser-Induced Breakdown Spectrometry for Depth Profiling of Coated Materials. *Applied Spectroscopy*. 2000, **54**(7), 1027–1031. ISSN 0003-7028. Available at: doi:10.1366/0003702001950526

[8] Harmon, RS., RE. Russo and RR. Hark. Applications of laser-induced breakdown spectroscopy for geochemical and environmental analysis: A comprehensive review. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2013, **87**, 11–26. ISSN 05848547. Available at: doi:10.1016/j.sab.2013.05.017

[9] Maurice, S., RC. Wiens, M. Saccoccio, B. Barraclough, O. Gasnault, O. Forni, N. Mangold, D. Baratoux, S. Bender, G. Berger, J. Bernardin, M. Berthé, N. Bridges, D. Blaney, M. Bouyé, P. Caïs, B. Clark, S. Clegg, A. Cousin, D. Cremers, A. Cros, L. DeFlores, C. Derycke, B. Dingler, G. Dromart, B. Dubois, M. Dupieux, E. Durand, L. D'Uston, C. Fabre, B. Faure, A. Gaboriaud, T. Gharsa, K. Herkenhoff, E. Kan, L. Kirkland, D. Kouach, J-L. Lacour, Y. Langevin, J. Lasue, S. Le Mouélic, M. Lescure, E. Lewin, D. Limonadi, G. Manhès, P. Mauchien, C. McKay, P-Y. Meslin, Y. Michel, E. Miller, HE. Newsom, G. Orttner, A. Paillet, L. Parès, Y. Parot, R. Pérez, P. Pinet, F. Poitrasson, B. Quertier, B. Sallé, C. Sotin, V. Sautter, H. Séran, JJ. Simmonds, J-B. Sirven, R. Stiglich, N. Striebig, J-J. Thocaven, MJ. Toplis and D. Vaniman. The ChemCam Instrument Suite on the Mars Science Laboratory

(MSL) Rover: Science Objectives and Mast Unit Description. *Space Science Reviews*. 2012, **170**(1–4), 95–166. ISSN 0038-6308. Available at: doi:10.1007/s11214-012-9912-2

[10] Nelson, T., R. Wiens, S. Clegg, R. Newell, S. Robinson, S. Storms, J. Michel, M. Caffrey, J. Deming, B. Sandoval, S. Maurice, P. Bernardi, P. Cais and F. Rull. The SuperCam Instrument for the Mars 2020 Rover. In: *2020 IEEE Aerospace Conference*. B.m.: IEEE, 2020, p. 1–12. ISBN 978-1-7281-2734-7. Available at: doi:10.1109/AERO47225.2020.9172661

[11] Hahn, DW. and N. Omenetto. Laser-Induced Breakdown Spectroscopy (LIBS), Part II: Review of Instrumental and Methodological Approaches to Material Analysis and Applications to Different Fields. *Applied Spectroscopy*. 2012, **66**(4), 347–419. ISSN 0003-7028. Available at: doi:10.1366/11-06574

[12] De Giacomo, A. and J. Hermann. Laser-induced plasma emission: from atomic to molecular spectra. *Journal of Physics D: Applied Physics*. 2017, **50**(18), 183002. ISSN 0022-3727. Available at: doi:10.1088/1361-6463/aa6585

[13] Capitelli, M., A. Casavola, G. Colonna and A. De Giacomo. Laser-induced plasma expansion: theoretical and experimental aspects. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2004, **59**(3), 271–289. ISSN 05848547. Available at: doi:10.1016/j.sab.2003.12.017

[14] Gornushkin, IB. and U. Panne. Radiative models of laser-induced plasma and pump-probe diagnostics relevant to laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2010, **65**(5), 345–359. ISSN 05848547. Available at: doi:10.1016/j.sab.2010.03.021

[15] Zhao, XZ., LJ. Shen, TX. Lu and K. Niemax. Spatial distributions of electron density in microplasmas produced by laser ablation of solids. *Applied Physics B Photophysics and Laser Chemistry*. 1992, **55**(4), 327–330. ISSN 0721-7269. Available at: doi:10.1007/BF00333075

[16] Aguilera, J. and C. Aragón. Multi-element Saha–Boltzmann and Boltzmann plots in laser-induced plasmas. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2007, **62**(4), 378–385. ISSN 0584-8547. Available at: doi:10.1016/J.SAB.2007.03.024

[17] Sankar, P., JJJ. Nivas, N. Smijesh, GK. Tiwari and R. Philip. Optical emission and dynamics of aluminum plasmas produced by ultrashort and short laser pulses. *Journal of Analytical Atomic Spectrometry*. 2017, **32**(6), 1177–1185. ISSN 0267-9477. Available at: doi:10.1039/C7JA00133A

[18] Tian, Y., EB. Sokolova, R. Zheng, Q. Ma, Y. Chen and J. Yu. Characteristics of the ablation plume induced on glasses for analysis purposes with laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2015, **114**, 7–14. ISSN 05848547. Available at: doi:10.1016/j.sab.2015.09.024

[19] Pan, Y., K. Tomita, K. Uchino, A. Sunahara and K. Nishihara. Time-resolved two-dimensional measurements of the electron density, electron temperature, and drift velocity of laser-produced carbon plasmas using the ion feature of collective laser Thomson scattering. *Applied Physics Express*. 2021, **14**(6), 066001. ISSN 1882-0778.

Available at: doi:10.35848/1882-0786/abfeca

[20] Casavola, AR., G. Colonna, A. De Giacomo, O. De Pascale and M. Capitelli. Experimental and theoretical investigation of laser-induced plasma of a titanium target. *Applied Optics*. 2003, **42**(30), 5963. ISSN 0003-6935. Available at: doi:10.1364/AO.42.005963

[21] Russo, RE. Laser Ablation. *Applied Spectroscopy*. 1995, **49**(9), 14A-28A. ISSN 0003-7028. Available at: doi:10.1366/0003702953965399

[22] Mao, XL., AC. Ciocan and RE. Russo. Preferential Vaporization during Laser Ablation Inductively Coupled Plasma Atomic Emission Spectroscopy. *Applied Spectroscopy*. 1998, **52**(7), 913–918. ISSN 0003-7028. Available at: doi:10.1366/0003702981944706

[23] Gornushkin, IB., JM. Anzano, LA. King, BW. Smith, N. Omenetto and JD. Winefordner. Curve of growth methodology applied to laser-induced plasma emission spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 1999, **54**(3–4), 491–503. ISSN 05848547. Available at: doi:10.1016/S0584-8547(99)00004-X

[24] Wagner, HG. C. Th. J. Alkemade, Tj. Hollander, W. Snelleman, P. J. Th. Zeegers: Metal Vapours in Flames. Pergamon Press 1982. Preis: DM 265,—. *Berichte der Bunsengesellschaft für physikalische Chemie*. 1983, **87**(11), 1104–1105. Available at: doi:https://doi.org/10.1002/bbpc.19830871140

[25] Tian, Y., B. Xue, J. Song, Y. Lu and R. Zheng. Non-gated laser-induced breakdown spectroscopy in bulk water by position-selective detection. *Applied Physics Letters*. 2015, **107**(11), 111107. ISSN 0003-6951. Available at: doi:10.1063/1.4931128

[26] Diaz, D. and DW. Hahn. Plasma chemistry produced during laser ablation of graphite in air, argon, helium and nitrogen. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2020, **166**, 105800. ISSN 05848547. Available at: doi:10.1016/j.sab.2020.105800

[27] Fu, Y-T., W-L. Gu, Z-Y. Hou, SA. Muhammed, T-Q. Li, Y. Wang and Z. Wang. Mechanism of signal uncertainty generation for laser-induced breakdown spectroscopy. *Frontiers of Physics*. 2021, **16**(2), 22502. ISSN 2095-0462. Available at: doi:10.1007/s11467-020-1006-0

[28] Tognoni, E. and G. Cristoforetti. Signal and noise in Laser Induced Breakdown Spectroscopy: An introductory review. *Optics and Laser Technology*. 2016, **79**(May), 164–172. ISSN 00303992. Available at: doi:10.1016/j.optlastec.2015.12.010

[29] Wang, Z., MS. Afgan, W. Gu, Y. Song, Y. Wang, Z. Hou, W. Song and Z. Li. Recent advances in laser-induced breakdown spectroscopy quantification: From fundamental understanding to data processing. *TrAC Trends in Analytical Chemistry*. 2021, **143**, 116385. ISSN 01659936. Available at: doi:10.1016/j.trac.2021.116385

[30] Zhang, T., H. Tang and H. Li. Chemometrics in laser-induced breakdown spectroscopy. *Journal of Chemometrics*. 2018, **32**(11), e2983. ISSN 08869383. Available at: doi:10.1002/cem.2983

[31] Howard, M. *Principles and practice of spectroscopic calibration*. New York: Wiley, 1991. ISBN 0471546143 9780471546146.

[32] Meloun, M., J. Militký, M. Hill and RG. Brereton. Crucial problems in regression

modelling and their solutions. *The Analyst*. 2002, **127**(4), 433–450. ISSN 00032654. Available at: doi:10.1039/b110779h

[33] Weiss, K., TM. Khoshgoftaar and D. Wang. A survey of transfer learning. *Journal of Big Data*. 2016, **3**(1), 9. ISSN 2196-1115. Available at: doi:10.1186/s40537-016-0043-6

[34] Ling, X., W. Dai, G-R. Xue, Q. Yang and Y. Yu. Spectral domain-transfer learning. In: *KDD*. 2008, p. 488–496. Available at: https://doi.org/10.1145/1401890.1401951

[35] Lee, S-I., V. Chatalbashev, D. Vickrey and D. Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In: *Proceedings of the 24th international conference on Machine learning - ICML '07*. New York, New York, USA: ACM Press, 2007, p. 489–496. ISBN 9781595937933. Available at: doi:10.1145/1273496.1273558

[36] Jebara, T. Multi-task feature and kernel selection for SVMs. In: *Twenty-first international conference on Machine learning - ICML '04*. New York, New York, USA: ACM Press, 2004, p. 55. ISBN 1581138285. Available at: doi:10.1145/1015330.1015426

[37] Jiang, J. and C. Zhai. Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007*. 2007, 264–271. Available at: https://ci.nii.ac.jp/naid/10026771973/en/

[38] Adams, N. Dataset Shift in Machine Learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2009, **173**(1), 274. Available at: doi:10.1111/j.1467-985X.2009.00624_10.x

[39] Vrbancic, G. and V. Podgorelec. Transfer Learning With Adaptive Fine-Tuning. *IEEE Access*. 2020, **8**, 196197–196211. ISSN 2169-3536. Available at: doi:10.1109/ACCESS.2020.3034343

[40] Guo, Y., H. Shi, A. Kumar, K. Grauman, T. Rosing and R. Feris. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[41] Glorot, X., A. Bordes and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: *ICML*. 2011.

[42] Tommasi, T. and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: *BMVC*. 2009.

[43] Tommasi, T., F. Orabona and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, p. 3081–3088.

[44] Shabbir, S., Y. Zhang, C. Sun, Z. Yue, W. Xu, L. Zou, F. Chen and J. Yu. Transfer learning improves the prediction performance of a LIBS model for metals with an irregular surface by effectively correcting the physical matrix effect. *Journal of Analytical Atomic Spectrometry*. 2021, **36**(7), 1441–1454. ISSN 0267-9477. Available at: doi:10.1039/D1JA00076D

[45] Yang, J., X. Li, H. Lu, J. Xu and H. Li. An LIBS quantitative analysis method for alloy steel at high temperature based on transfer learning. *Journal of Analytical Atomic*

Spectrometry. 2018, **33**(7), 1184–1195. ISSN 0267-9477. Available at: doi:10.1039/C8JA00069G

[46] Kaneko, H., S. Kono, A. Nojima and T. Kambayashi. Transfer learning and wavelength selection method in NIR spectroscopy to predict glucose and lactate concentrations in culture media using VIP-Boruta. *Analytical Science Advances*. 2021, ansa.202000177. ISSN 2628-5452. Available at: doi:10.1002/ansa.202000177

[47] Sun, C., W. Xu, Y. Tan, Y. Zhang, Z. Yue, S. Shabbir, M. Wu, L. Zou, F. Chen and J. Yu. From Machine Learning to Transfer Learning in Laser-Induced Breakdown Spectroscopy: the Case of Rock Analysis for Mars Exploration. 2021. Available at: http://arxiv.org/abs/2102.03768

[48] Chang, F., H. Lu, H. Sun and J. Yang. Assessment of the performance of quantitative feature-based transfer learning LIBS analysis of chromium in high temperature alloy steel samples. *Journal of Analytical Atomic Spectrometry*. 2020, **35**(11), 2639–2648. ISSN 0267-9477. Available at: doi:10.1039/D0JA00334D

[49] Yun, Y-H., H-D. Li, B-C. Deng and D-S. Cao. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends in Analytical Chemistry*. 2019, **113**, 102–115. ISSN 01659936. Available at: doi:10.1016/j.trac.2019.01.018

[50] Maaten, L Van Der., E. Postma and J. Van Den Herik. Dimensionality reduction: A comparative review. *J Mach Learn Res*. 2009, **10**(66–71), 13.

[51] Pořízka, P., J. Klus, E. Képeš, D. Prochazka, DW. Hahn and J. Kaiser. On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review. *Spectrochim. Acta - Part B At. Spectrosc.* 2018, **148**, 65–82. ISSN 05848547. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0584854718301526

[52] Joliffe, IT. *Principal Component Analysis*. New York: Springer-Verlag, 2002. Springer Series in Statistics. ISBN 0-387-95442-2. Available at: doi:10.1007/b98835

[53] Vrábel, J., E. Képeš, L. Duponchel, V. Motto-Ros, C. Fabre, S. Connemann, F. Schreckenberg, P. Prasse, D. Riebe, R. Junjuri, MK. Gundawar, X. Tan, P. Pořízka and J. Kaiser. Classification of challenging Laser-Induced Breakdown Spectroscopy soil sample data - EMSLIBS contest. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2020, **169**, 105872. ISSN 05848547. Available at: doi:10.1016/j.sab.2020.105872

[54] Boucher, T., C. Carey, MD. Dyar, S. Mahadevan, S. Clegg and R. Wiens. Manifold preprocessing for laser-induced breakdown spectroscopy under Mars conditions. *Journal of Chemometrics*. 2015, **29**(9), 484–491. ISSN 08869383. Available at: doi:10.1002/cem.2727

[55] Schölkopf, B., A. Smola and K-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*. 1998, **10**(5), 1299–1319. ISSN 0899-7667. Available at: doi:10.1162/089976698300017467

[56] Schölkopf, B., R. Herbrich and AJ. Smola. A Generalized Representer Theorem. In: . 2001, p. 416–426. Available at: doi:10.1007/3-540-44581-1_27

[57] Belkin, M. and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*. 2003, **15**(6), 1373–1396. ISSN 0899-7667.

Available at: doi:10.1162/089976603321780317

[58] Belkin, M. and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Cambridge, MA, USA: MIT Press, 2001, p. 585–591. NIPS'01.

[59] Chen, J. and Y. Liu. Locally linear embedding: a survey. *Artificial Intelligence Review*. 2011, **36**(1), 29–48. ISSN 0269-2821. Available at: doi:10.1007/s10462-010-9200-z

[60] Roweis, ST. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*. 2000, **290**(5500), 2323–2326. ISSN 00368075. Available at: doi:10.1126/science.290.5500.2323

[61] Tenenbaum, JB. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*. 2000, **290**(5500), 2319–2323. ISSN 00368075. Available at: doi:10.1126/science.290.5500.2319

[62] Kamada, T. and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*. 1989, **31**(1), 7–15. ISSN 00200190. Available at: doi:10.1016/0020-0190(89)90102-6

[63] Motto-Ros, V., D. Syvilay, L. Bassel, E. Negre, F. Trichard, F. Pelascini, J. El Haddad, A. Harhira, S. Moncayo, J. Picard, D. Devismes and B. Bousquet. Critical aspects of data analysis for quantification in laser-induced breakdown spectroscopy. *Spectrochimica Acta - Part B Atomic Spectroscopy*. 2018, **140**, 54–64. ISSN 05848547. Available at: doi:10.1016/j.sab.2017.12.004

[64] Guezenoc, J., L. Bassel, A. Gallet-Budynek and B. Bousquet. Variables selection: A critical issue for quantitative laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2017, **134**, 6–10. ISSN 05848547. Available at: doi:10.1016/j.sab.2017.05.009

[65] Gonzaga, FB. and C. Pasquini. A Complementary Metal Oxide Semiconductor sensor array based detection system for Laser Induced Breakdown Spectroscopy: Evaluation of calibration strategies and application for manganese determination in steel. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2008, **63**(1), 56–63. ISSN 05848547. Available at: doi:10.1016/j.sab.2007.11.005

[66] Barbieri Gonzaga, F. and C. Pasquini. A compact and low cost laser induced breakdown spectroscopic system: Application for simultaneous determination of chromium and nickel in steel using multivariate calibration. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2012, **69**, 20–24. ISSN 05848547. Available at: doi:10.1016/j.sab.2012.02.007

[67] Bousquet, B., JB. Sirven and L. Canioni. Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples. *Spectrochimica Acta - Part B Atomic Spectroscopy*. 2007, **62**(12), 1582–1589. ISSN 05848547. Available at: doi:10.1016/j.sab.2007.10.018

[68] Pontes, MJC., J. Cortez, RKH. Galvão, C. Pasquini, MCU. Araújo, RM. Coelho, MK. Chiba, MF. De Abreu and BE. Madari. Classification of Brazilian soils by using LIBS and variable

selection in the wavelet domain. *Analytica Chimica Acta*. 2009, **642**(1–2), 12–18. ISSN 00032670. Available at: doi:10.1016/j.aca.2009.03.001

[69] Mehta, P., M. Bukov, C-H. Wang, AGRR. Day, C. Richardson, CK. Fisher and DJ. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*. 2019, **810**, 1–124. ISSN 03701573. Available at: doi:10.1016/j.physrep.2019.03.001

[70] Sjöberg, J. and L. Ljung. Overtraining, Regularization, and Searching for Minimum in Neural Networks. *IFAC Proceedings Volumes*. 1992, **25**(14), 73–78. ISSN 14746670. Available at: doi:10.1016/S1474-6670(17)50715-6

[71] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996, **58**(1), 267–288. ISSN 00359246. Available at: doi:10.1111/j.2517-6161.1996.tb02080.x

[72] Brereton, RG. Introduction to multivariate calibration in analytical chemistry. *The Analyst*. 2000, **125**(11), 2125–2154. ISSN 00032654. Available at: doi:10.1039/b003805i

[73] Webb, GI., C. Sammut, C. Perlich, T. Horváth, S. Wrobel, KB. Korb, WS. Noble, C. Leslie, MG. Lagoudakis, N. Quadrianto, WL. Buntine, N. Quadrianto, WL. Buntine, L. Getoor, G. Namata, L. Getoor, J. Han, Xin Jin, J-A. Ting, S. Vijayakumar, S. Schaal and L. De Raedt. Leave-One-Out Cross-Validation. In: *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2011, p. 600–601. Available at: doi:10.1007/978-0-387-30164-8_469

[74] Rand, WM. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*. 1971, **66**(336), 846–850. ISSN 0162-1459. Available at: doi:10.1080/01621459.1971.10482356

[75] Rousseeuw, PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987, **20**, 53–65. ISSN 03770427. Available at: doi:10.1016/0377-0427(87)90125-7

[76] Bezdek, JC. and NR. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*. 1998, **28**(3), 301–315. ISSN 10834419. Available at: doi:10.1109/3477.678624

[77] Dyar, MD., ML. Carmosino, EA. Breves, MV. Ozanne, SM. Clegg and RC. Wiens. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2012, **70**, 51–67. ISSN 05848547. Available at: doi:10.1016/j.sab.2012.04.011

[78] Segev, N., M. Harel, S. Mannor, K. Crammer and R. El-Yaniv. Learn on Source, Refine on Target:A Model Transfer Learning Framework with Random Forests. 2015. Available at: doi:10.1109/TPAMI.2016.2618118

[79] Schetinin, V., JE. Fieldsend, D. Partridge, TJ. Coats, WJ. Krzanowski, RM. Everson, TC. Bailey and A. Hernandez. Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications. *IEEE Transactions on Information Technology in Biomedicine*. 2007, **11**(3), 312–319. ISSN 1089-7771. Available

at: doi:10.1109/TITB.2006.880553

[80] Ranzato, F. and M. Zanella. Abstract Interpretation of Decision Tree Ensemble Classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, **34**(04), 5478–5486. ISSN 2374-3468. Available at: doi:10.1609/aaai.v34i04.5998

[81] Boser, BE., IM. Guyon and VN. Vapnik. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. New York, New York, USA: ACM Press, 1992, p. 144–152. ISBN 089791497X. Available at: doi:10.1145/130385.130401

[82] Cortes, C. and V. Vapnik. Support-vector networks. *Machine Learning*. 1995, **20**(3), 273–297. ISSN 0885-6125. Available at: doi:10.1007/BF00994018

[83] Anthony, G., H. Gregg and M. Tshilidzi. Image Classification Using SVMs: One-against-One Vs One-against-All. 2007. Available at: http://arxiv.org/abs/0711.2914

[84] Schölkopf, B. The kernel trick for distances. *Advances in neural information processing systems*. 2001, 301–307.

[85] Lixin Duan, Dong Xu and IW. Tsang. Domain Adaptation From Multiple Sources: A Domain-Dependent Regularization Approach. *IEEE Transactions on Neural Networks and Learning Systems*. 2012, **23**(3), 504–518. ISSN 2162-237X. Available at: doi:10.1109/TNNLS.2011.2178556

[86] Barrett, HH. III The Radon Transform and Its Applications. In: . 1984, p. 217–286. Available at: doi:10.1016/S0079-6638(08)70123-9

[87] Gornushkin, IB., S. Merk, A. Demidov, U. Panne, SV. Shabanov, BW. Smith and N. Omenetto. Tomography of single and double pulse laser-induced plasma using Radon transform technique. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2012, **76**, 203–213. ISSN 05848547. Available at: doi:10.1016/j.sab.2012.06.033

[88] Kak, AC. and M. Slaney. *Principles of Computerized Tomographic Imaging*. B.m.: Society for Industrial and Applied Mathematics, 2001. ISBN 978-0-89871-494-4. Available at: doi:10.1137/1.9780898719277

[89] Képeš, E., I. Gornushkin, P. Pořízka and J. Kaiser. Spatiotemporal spectroscopic characterization of plasmas induced by non-orthogonal laser ablation. *The Analyst*. 2021, **146**(3), 920–929. ISSN 0003-2654. Available at: doi:10.1039/D0AN01996H

[90] Képeš, E., I. Gornushkin, P. Pořízka and J. Kaiser. Spatiotemporal spectroscopic characterization of plasmas induced by non-orthogonal laser ablation. *Analyst*. 2021, **146**(3). ISSN 13645528. Available at: doi:10.1039/d0an01996h

[91] Képeš, E., I. Gornushkin, P. Pořízka and J. Kaiser. Tomography of double-pulse laser-induced plasmas in the orthogonal geometry. *Analytica Chimica Acta*. 2020, **1135**, 1–11. ISSN 00032670. Available at: doi:10.1016/j.aca.2020.06.078

[92] Breves, EA., K. Lepore, MD. Dyar, SC. Bender, RL. Tokar and T. Boucher. Laser-induced breakdown spectra of rock powders at variable ablation and collection angles under Mars-analog conditions. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2017, **137**, 46–58. ISSN 05848547. Available at: doi:10.1016/j.sab.2017.09.002

[93] Klus, J., P. Pořízka, D. Prochazka, J. Novotný, K. Novotný and J. Kaiser. Effect of experimental parameters and resulting analytical signal statistics in laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2016, **126**, 6–10. ISSN 05848547. Available at: doi:10.1016/j.sab.2016.10.002

[94] Képeš, E., P. Pořízka and J. Kaiser. On the application of bootstrapping to laser-induced breakdown spectroscopy data. *Journal of Analytical Atomic Spectrometry*. 2019, **34**(12), 2411–2419. ISSN 0267-9477. Available at: doi:10.1039/C9JA00304E

[95] Képeš, E., P. Porízka and J. Kaiser. On the application of bootstrapping to laser-induced breakdown spectroscopy data. *Journal of Analytical Atomic Spectrometry*. 2019, **34**(12). ISSN 13645544. Available at: doi:10.1039/c9ja00304e

[96] Képeš, E., P. Pořízka, J. Klus, P. Modlitbová and J. Kaiser. Influence of baseline subtraction on laser-induced breakdown spectroscopic data. *Journal of Analytical Atomic Spectrometry*. 2018, **33**(12), 2107–2115. ISSN 0267-9477. Available at: doi:10.1039/C8JA00209F

[97] Képeš, E., J. Vrábel, P. Pořízka and J. Kaiser. Addressing the sparsity of laser-induced breakdown spectroscopy data with randomized sparse principal component analysis. *Journal of Analytical Atomic Spectrometry*. 2021, **36**(7), 1410–1421. ISSN 0267-9477. Available at: doi:10.1039/D1JA00067E

[98] Erichson, NB., S. Voronin, SL. Brunton and JN. Kutz. Randomized Matrix Decompositions Using R. *Journal of Statistical Software*. 2016, **89**(11). ISSN 1548-7660. Available at: doi:10.18637/jss.v089.i11

[99] Wiens, RC., S. Maurice, J. Lasue, O. Forni, RB. Anderson, S. Clegg, S. Bender, D. Blaney, BL. Barraclough, A. Cousin, L. Deflores, D. Delapp, MD. Dyar, C. Fabre, O. Gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P-Y. Meslin, H. Newsom, A. Ollila, R. Perez, RL. Tokar and D. Vaniman. Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover. *Spectrochimica Acta Part B: Atomic Spectroscopy*. 2013, **82**, 1–27. ISSN 05848547. Available at: doi:10.1016/j.sab.2013.02.003

[100] McInnes, L., J. Healy and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Available at: http://arxiv.org/abs/1802.03426

[101] Képeš, E., J. Vrábel, S. Střítežská, P. Pořízka and J. Kaiser. Benchmark classification dataset for laser-induced breakdown spectroscopy. *Scientific Data*. 2020, **7**(1), 53. ISSN 2052-4463. Available at: doi:10.1038/s41597-020-0396-8