Master Thesis

Empirical simulation of laser-induced breakdown spectroscopy spectra

carried out at
Institute of Chemical Technologies and Analytics,
Technical University of Vienna

under the supervision of
Ao.Univ.Prof. Mag.rer.nat. Dr.rer.nat. Johann Lohninger
Univ.Ass.in Dipl.-Ing.in Zuzana Gajarska

by
Marlene Nadvornik

January 2024

Marlene Nadvornik

Abstract

The combination of laser-induced breakdown spectroscopy (LIBS) and machine learning is raising a lot of interest due to a wide range of potential applications including identification of polymers, classification of soils and diagnosis of cancer. However, a general challenge of classification problems is robustness against unknown samples. One possibility to overcome this issue while cutting down production costs is to increase the number of samples and the variety of sample compositions by an in-silico production of spectra.

As the conventional approach relying on theoretical modelling of the underlying physicochemical phenomena fails to capture the variations of the spectra arising from factors such as sample inhomogeneity and other complex sample-laser interactions, this thesis investigates an alternative approach based on empirical modelling. Throughout the work, polynomial models of three different materials (low-alloyed steel, aluminium alloy, and borosilicate glass) reflecting the changes in laser energy and gate delay were developed and discussed.

Each of the three spectral constituents (the noise, the baseline, and the emission lines) was modelled separately. The noise exhibited a heteroscedastic character. Nevertheless, considering the insignificant effect of noise on the final model, the modelling effort was reduced by assuming homoscedastic noise irrespective of gate delay and laser energy. For the baseline and emission lines, bilinear polynomial models, accounting for the variation in laser energy and gate delay, were constrained to a maximum degree of five and two, respectively. The entire development process was carried out using Python.

While the final models could roughly simulate the measurement conditions of the used materials, a more realistic simulation would require an in-depth tuning of the processing steps. Yet, the existing models are valuable for explorative tasks. Therefore, a custom-made program coded in Delphi-Object Pascal was provided.

Kurzfassung

Die Kombination von laserinduzierter Plasmaspektroskopie (LIBS) und maschinellem Lernen erfährt steigendes Interesse durch eine Vielzahl an potenziellen Anwendungen wie beispielsweise die Identifikation von Polymeren, die Klassifizierung von Bodenproben oder die Diagnose von Krebs. Jedoch ist die Robustheit gegen unbekannte Proben eine generelle Herausforderung von Klassifikationsmodellen. Eine Möglichkeit dieses Problem zu überwinden und gleichzeitig Kosten zu senken ist sowohl die Erhöhung der Probenanzahl als auch der Vielfältigkeit der Probenzusammensetzungen mittels künstlich simulierter Spektren.

Der konventionelle Ansatz, basierend auf der Modellierung der zugrunde liegenden physikochemischen Prozesse, zeigt Defizite in Bezug auf Veränderungen im Spektrum, welche durch inhomogene Probenzusammensetzung oder komplexe Laser-Probe-Interaktionen ausgelöst werden. In dieser Arbeit wird dafür ein neuer Ansatz basierend auf empirischen Simulationen entwickelt. Hierbei wurden Polynommodelle von drei unterschiedlichen Materialien (niedrig-legierter Stahl, Aluminiumlegierung, Borosilikatglas) in Abhängigkeit von Laserenergie und Gate-Verzögerung entwickelt und disktutiert.

Jeder der drei Bestandteile der Spektren (Rauschen, Basislinie und Emissionslinien) wurde getrennt modelliert. Das Rauschen wies heteroskedastischen Charakter auf. Dennoch wurde für das finale Modell homoskedastisches Rauschen unabhängig von Laserenergie und Gate-Verzögerung angenommen, um den Modellierungsaufwand zu reduzieren. Dies war angesichts der unbedeutenden Auswirkungen des Rauschens auf die endgültige Simulation vertretbar. Für die Modelle der Basislinie und der Elementlinien wurden die bilinearen Polynommodelle durch einen maximalen Grad von fünf bzw. zwei begrenzt. Der komplette Entwicklungsprozess wurde in Python programmiert.

Die finalen Modelle waren grob in der Lage die Messbedingungen der verwendeten Materialien zu simulieren. Für eine realistischer Simulation wäre eine weiterführende gründliche Anpassung der Modellierungsschritte notwendig. Dennoch können die bereits bestehenden Modelle für explorative Aufgaben verwendet werden. Dafür, wurde auch ein spezielles Programm in Delphi/Object Pascal programmiert.

Acknowledgements

First of all, I would like to express my gratitude to Hans who not only enabled me to do this master thesis but who supported me throughout this thesis in the best way a student could think of. I do not take the trust to perform a pure coding-based master thesis as a chemistry student with only limited prior coding experience for granted. By giving me the freedom to work completely independently, I was able to deepen my Python knowledge significantly along the way. Even more important, introducing me to the world of software development in cooperation with the Epina Softwareentwicklungs- und Vertriegs-GmbH expanded my mental landscape substantially and is something I will cherish throughout my future career.

Given the fact that our research group was only made up of three people - including me, I always referred to the group as a little family. Anyhow, I would not do so if it was not for Zuzana. She made me feel welcomed from the first moment on and her support reached way beyond the extent of the master thesis. The thesis could not have been rounded off better than the experiences at the Europython conference together.

Furthermore, I would like to thank my colleagues Sebastian and Yikai who might not have even noticed how much they helped me to enhance my coding skills by just working with me together on projects for a machine learning course.

Last but not least, I thank my family and friends for the support you have given me not only throughout this thesis but for the entire degree. Especially without my parents, none of my educational milestones would have been possible.

Content

List of acronyms

Abbreviations

CCD	charge-coupled device
DF	detail factor
DWR	detailed wavelength range
EMCCD	electron-multiplying CCD
EPW	endpoint-weight
LE	laser energy
LIBS	laser-induced breakdown spectroscopy
LS	line shift
IB	inverse bremsstrahlung
GD	gate delay
MAD	median absolute deviation
MCP	micro-channel plate
MLR	multiple linear regression
MSE	mean square error
MSR	regression mean square
Nd:YAG	neodymium-doped yttrium aluminium garnet
PCC	Pearson correlation coefficient
PDA	photo-diode array
PHW	peak half width
PMT	photomultiplier tube
REVD	reverse direction
RSD	relative standard deviation
RSS	residual sum of squares
TSS	total sum of squares
UV	ultra-violet
WW_{med}	window width of the median filter
WW_{min}	window width of the minima filter
WW_{std}	window width of the standard deviation filter
YAG	yttrium aluminium garnet

Symbols

A	peak height
a	coefficient
d_r	rank difference
h	Planck constant
n	number of observations
k	number of independent variables
I	intensity
$\tilde{I}_c$	median intensity of a measurement condition
p	polynomial degree
R^2	coefficient of determination
s_d	standard deviation
s_e	standard error
γ	curve width
$\tilde{\delta}$	median sign
θ	shape coefficient
λ	wavelength
λ_0	center wavelength
ν	frequency
ρ_S	Spearman's rank correlation coefficient
σ	standard deviation
ϕ	power transformation parameter

1 Introduction

1.1 Purpose

State of the art

Laser-induced breakdown spectroscopy (LIBS) is increasingly combined with machine learning algorithms, resulting in a broader range of applications. Some examples are the identification of polymer types (Gajarska et al.[1]), the classification of soils (Pontes et al.[2]) or the diagnosis of cancer via the LIBS analysis of human plasma (Yue et al.[3]). Nevertheless, the lack of robustness against unknown samples remains to be a general problem. Further, for LIBS, poor reproducibility of measurements across different LIBS instruments as well as measurement conditions increases the difficulty of universal application of classification models. To overcome this issue, training models with more diverse datasets is necessary. Though, one possibility while cutting down production cost is to increase the variety and number of samples by simulating spectra.

Context of the master thesis

The conventional approach for simulation of LIBS spectra relies on the theoretical modelling of complex physicochemical phenomena. Nevertheless, resulting models are often limited due to inherent assumptions in the theoretical framework. Thus, this thesis aims to present a novel approach for simulating LIBS spectra based on empirical modelling of three different materials (low-alloyed steel, aluminum alloy, and borosilicate glass) under various experimental conditions, including variations in laser energy and gate delay. Through this approach, the thesis aims to overcome limitations associated with assumptions made in theoretical models, leading to more accurate and realistic representations of LIBS spectra.

Moreover, a custom-built user interface should be provided to use the generated models for explorative purposes.

1.2 Laser-induced breakdown spectroscopy

In the 1860s, it was found that the emitted light spectra of elements can be used as their fingerprint. Since than, several analytical techniques utilizing the emission of electromagnetic radiation of excited atoms, ions and molecules were developed. LIBS being one of them.[4] Since its development in the 1980s, LIBS has evolved into a powerful analytical tool for providing real-time measurements of constituents in almost any kind of material.
The principle of LIBS is rather simple, although the physical processes involved in the laser-matter interaction are complex. It is based on focusing a pulsed laser down to a target (solid, liquid, gas) to ablate (given the sample is solid) and vaporize a small amount of sample material. Within the generated plasma, optical emission by excited atomic, ionic and molecular species takes place. A fraction of the emitted light is then spectrally analysed and can be evaluated for qualitative and quantitative tasks.[5] A detailed explanation of the instrumental set-up and plasma processes follows in section 1.2.1 and paragraph 1.2.2.1.
Further perks of LIBS are the absence of sample preparation, the ability to perform in situ analysis as well as the quasi-non-destructive and micro-analysis character of measurements.[5]

1.2.1 Instrumental set-up

The main components of a LIBS instrument are: a pulsed laser that generates the optical pulses used to form the microplasma, the lens that focuses the laser pulse on the target, a target holder, the light collection system that collects and directs the light to the detection system, a spectrometer to disperse the light, a detector and control unit with a computer.[6] A schematic representation of a LIBS instrument is shown in Figure 1.

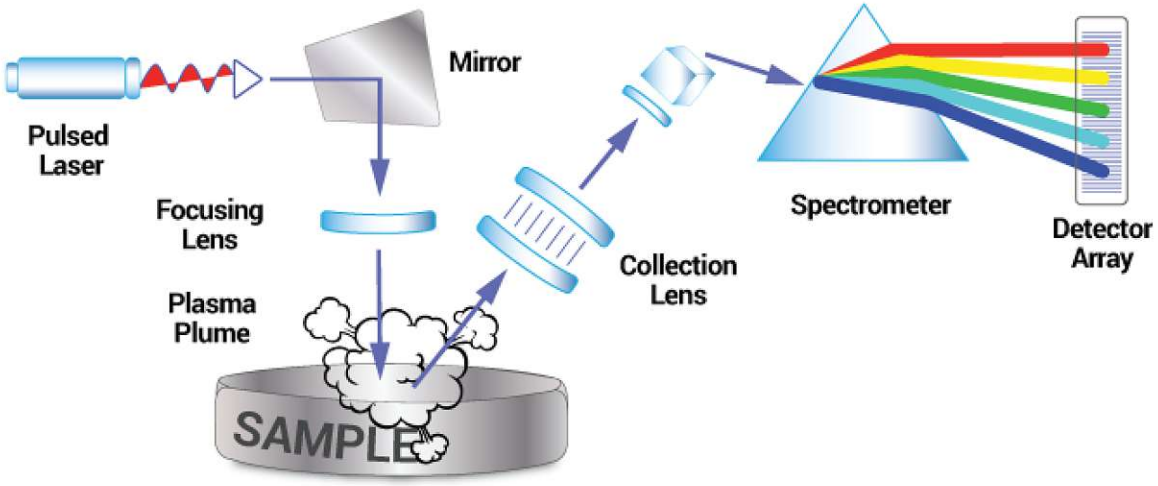


Figure 1: Schematic representation of a LIBS instrument.[7]

Laser

The discussion of lasers here is limited to neodymium-doped yttrium aluminium garnet (Nd:YAG) lasers since this type was used for measuring the data of this thesis.
In lasers, the interaction of photons with atoms can be divided into two processes: absorption of radiation and stimulated emission.[8] For the first one, a flashlamp or diode stack is fired to produce excitation in the lasing material. A small fraction of this pumping light is absorbed by ions doped into the lasing material. For a Nd:YAG laser, Nd^{3+} ions are doped into yttrium aluminium garnet (YAG) crystal matrix. This way, electrons in the ground state are pumped into a higher energy level. If the pumping excitation is sufficiently strong, a "population inversion" is established in which the upper level is more populated than the lower level. Now, photons passing through the lasing medium with the same frequency as the lasing transition will induce the decay of electrons to the ground state. This process is known as stimulated emission. If the medium is surrounded by a resonant cavity, the emitted light is directed back and significant amplification of light at the wavelength of the lasing transition can be achieved. To achieve high-power laser pulses an electro-optic Q-switch shutter positioned in the cavity is necessary. The Q-switch prevents photons at the transition wavelength from stimulating emission, leading to a drastic population inversion. After a certain time, the Q-switch becomes transparent and the laser is suddenly triggered resulting in a high-power pulse of short duration. A typical Q-switched pulse is in the length of 5-10 ns.[6,9]
Additionally, the following laser properties are a specific matter of interest for LIBS:

- Wavelength
  Some wavelengths can be better absorbed by a material than others. The fundamental frequency of a Nd:YAG laser is 1064 nm.
- Pulse energy
  Too low pulse energies prevent sufficient sample ablation and vaporization resulting in poor emission signal. Typical pulse energies range from 10 to 500 mJ.
- Focused pulse power density (irradiance)
  The irradiance is determined by pulse energy and pulse width. The shorter the pulse width, the greater the power density delivered onto the target.
- Spatial beam quality
  The spatial quality determines the minimum spot size to which the laser beam can be focused. For beams of poor quality, focusing on a sufficiently small spot might not always be possible.[6]

Optical System

To focus the laser pulse to a sufficiently small spot, lenses and mirrors can be used. Normally, a single lens is enough but multi-lens systems might be required in systems with adjustable focus. Likewise, a lens or mirror system is used to collect the plasma light. Albeit, lens systems are prone to exhibiting chromatic aberrations due to the dependence of the refractive index of optical materials on wavelength. So, the focal position of the lens system depends on the wavelength as well. Anyhow, fibre optic cables have prevailed to collect the plasma light. In particular, they allow applications where remote probing is necessary. The fibre optic cables are typically made up of fused silica with varying OH content depending on the demanded spectral range and the core diameters range from 50 μm to 1 mm.[6]

Spectrometer

To retrieve information from the laser plasma, the separate measurement of specific wavelengths is necessary. The applied architectures range from simple spectral line filters for a single wavelength to sophisticated designs such as an echelle spectrograph or grating spectrographs. Among grating spectrographs, the Czerny-Turner architecture is the most prominent variant. Here, the light from the plasma passes through an entrance slit before it is collimated at the first mirror and directed onto the grating. According to the wavelength, the light is reflected off the grating at different angles. The reflected line is then projected as a spectrum on a focal plane with a second mirror. It can then be recorded with a photo-diode array (PDA) or a charge-coupled device (CCD) array detector. However, the spectral range recorded simultaneously is limited by the width of the focal plane and the size of the detector array. Moreover, the Czerny-Turner design does not utilize the two-dimensionality of the CCD in contrary to echelle spectrographs.[6] In echelle spectrographs, the light is dispersed in two orthogonal directions using two dispersion stages. The two dispersive units can be gratings or prisms or a combination of them. The result is a display of the spectral data as a 2D pattern which can be ideally captured with a two-dimensional CCD camera. Consequently, echelle spectrographs can provide high spectral resolution as well as large spectral bandpass which makes them perfectly suitable for LIBS. Furthermore, echelle spectrographs can be incorporated into portable LIBS systems due to the absence of moving parts.[10] An echelle spectrograph was also used for the spectra processed within this thesis.

Detector

PDAs and CCDs are the most common detectors for Czerny-Turner and echelle spectrographs due to their capability of providing simultaneous acquisition of emission lines along a wide range of wavelengths. A photo-diode array is a one-dimensional array of discrete photo-diodes on an integrated circuit. Photodiodes itself are p-n junctions like conventional semiconductor diodes. When a photon strikes the diode in the junction's depletion region, an electron-hole pair is generated which produces a photocurrent whose intensity is proportional to the number of incident photons. In comparison, a CCD is an integrated circuit etched onto a silicon surface which forms light-sensitive elements named pixels. Here, free electrons generated in each pixel by incident photons are collected and stored in the semiconductor region underneath the pixel.[5] For bidimensional arrays, the CCD is read out by repeatedly shifting all the rows vertically downward into the shift register, which is then read out sequentially. Obviously, CCDs are preferred for echelle spectrographs due to their two-dimensional character. Yet, for 1D requirements as in Czerny-Turner spectrographs, PDA is the option of choice for single-shot measurements if the signal levels are near the saturation level.[6]

Gating

For LIBS detectors, fast gating capability of the sensors is of fundamental importance since it allows to control the time interval between the application of the laser pulse and the beginning of the signal detection (= gate delay).[5] As optimal gate delays are within the microsecond timescale and readout times of CCD arrays are on the order of a millisecond, a shutter is necessary. A typical shutter applied for array detectors is micro-channel plates (MCPs). A MCP consists of an array of parallel glass tubes which act like photomultiplier tubes. The gating itself is provided by the application of a high voltage between the front and back side of the MCP facilitating a very rapid permission or prohibition of light. The combination of a MCP and a CCD is also known as intensified CCD or ICCD.

Nevertheless, newer types of CCDs are electron-multiplying CCDs (EMCCDs). Here, an electron-multiplying gain register is placed between the CCD shift register and the output resulting in amplification factors of 10^3 even at high speeds. These detectors are especially suited when high readout rates are required. An EMCCD was also used to record the provided spectra within this master thesis.[6]

1.2.2 Plasma

1.2.2.1 Lifecycle

In general, a plasma is a local assembly of atoms, ions, molecules and free electrons in which the charged species often act collectively. Overall, the charge of a plasma is quasi-neutral. One of the major characteristics of a plasma is its degree of ionization. Here, plasmas are assigned as weakly ionized when the ratio of electrons to other species is less than 10 % and LIBS falls into this category. Further properties of plasma are the plasma temperature and the electron density.[6]

Moreover, the lifespan of a LIBS plasma can be separated into plasma ignition, plasma expansion and particle ejection, which will be discussed below.[11] A detailed discussion of these steps follows and a summary of the plasma processes can be seen in Figure 2.

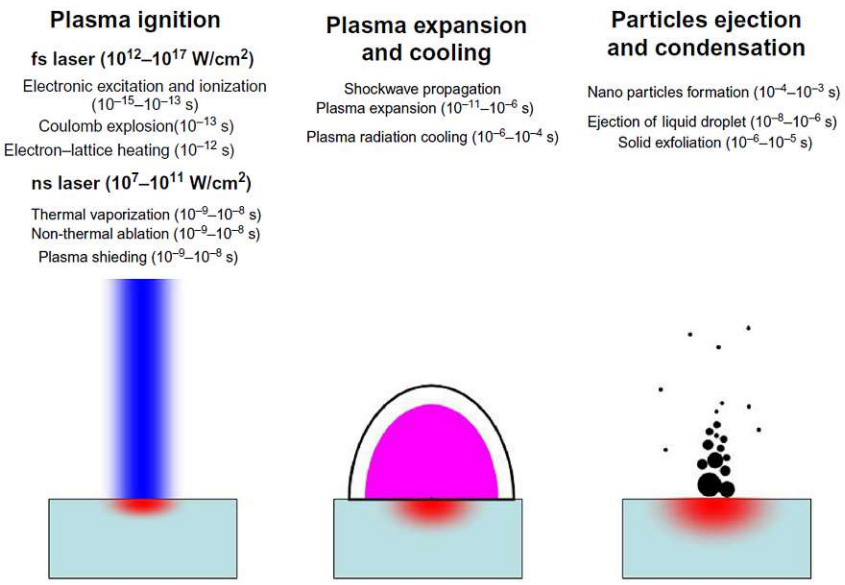


Figure 2: Summary of plasma processes in LIBS and various mechanisms occurring during each process[11]

Plasma ignition

Plasma ignition and its accompanying processes of bond breaking and plasma shielding are highly dependent on laser irradiance and pulse duration. Since nanosecond laser pulses were used for the measurement of the provided spectra of this thesis, the following discussion is limited to nanosecond pulse durations.
Depending on the laser irradiance thermal ($< 10^8$ W/cm^2) or non-thermal ($> 10^{11}$W/cm^2) processes are dominant for ablating the sample. If thermal vaporization is dominant, well-defined phase transitions take place as the surface temperature increases (solid → liquid → vapour → plasma). Concerning non-thermal processes, Coulomb explosion is the main bond-breaking mechanism. After vaporization, two mechanisms for electron generation are possible. The favoured mechanism depends on the laser wavelength (see section 1.2.4). The first mechanism is

inverse bremsstrahlung (IB), which involves the absorption of a photon by free electrons during a collision with heavy particles. The second mechanism is known as (multi)photon ionization where atoms or molecules absorb a sufficient number of photons to get ionized (Equation 1).[11,6]

$$\mathrm{M} + mh\nu \rightarrow \mathrm{M}^{+} + \mathrm{e}^{-} \tag{1}$$

atom or molecule (M); number of photons (m); Planck constant (h); frequency (ν); electron (e)

Afterwards, a fraction of the liberated electrons will have sufficient energy to interact with other atoms and molecules, as seen in Equation 2, resulting in an ionization cascade.

$$\mathrm{M} + \mathrm{e}^{-} \rightarrow \mathrm{M}^{+} + 2\mathrm{e}^{-} \tag{2}$$

atom or molecule (M); electron (e)

Plasma formation takes place if the laser intensity exceeds a threshold value. The threshold intensity depends on the thermal properties of the sample material.
Due to the relatively long pulse duration compared to femto- and picosecond lasers, vaporization and absorption are only present in an initial fraction of the pulse. For the remaining time, the laser pulse energy is partially absorbed in the vapour and the expanding plasma plume before reaching the target surface (plasma shielding). Here, absorption primarily takes place by IB if the plasma density is below the critical density, and the plasma expands normally to the target at supersonic speed. While expanding, the plasma compresses the surrounding gas along the establishment of shock waves and transfers energy to the ambient gas by a combination of thermal conduction, radiative transfer and heating via the shock wave. Moreover, the plasma temperature decreases as it expands away from the target. Plasma shielding will also influence the vapour conversion rate of the solid and the properties of the vapour.[11]

Plasma expansion

As the laser pulse stops, the evolved plasma plume will continue to expand into the surrounding atmosphere or vacuum. The plasma density and temperature will continue to change. Once the plasma pressure equals the ambient pressure, the expansion stops. This step is mainly dependent on the initial plasma properties and the expansion medium.[11]

Particle ejection

As the temperature of the plasma falls below the boiling point temperature of the material, nano-sized particles will be formed from the condensation of the vapour. Further, liquid ejection of particles may occur by high-pressure gradient forces within the expanding vapour plume. Additionally, solid sample exfoliation might be visible due to the thermal stress gradients of the fast heating process.[11]

1.2.2.2 Spectral emission

Continuum as well as line emission contribute to the spectral composition of a plasma.[11] The temporal appearance of these signals is discussed in section 1.2.4.

Continuum emission

Free-free and free-bound transition result in continuum radiation. Free-free transitions are also known as bremsstrahlung and occur due to the interaction of electrons with positively charged ions. If the plasma density is below the critical density IB occurs as part of plasma shielding instead of bremsstrahlung emission. The second source of bremsstrahlung transition results from recombination radiation in which a free electron is captured by an ion and the excess kinetic energy of the electron is emitted as a photon.[11]

Line emission

The observed intensity of a spectral line depends on the amount of element, the transition probability (an intrinsic property of the species) and on the environmental conditions in the plasma. So, to observe a line, the plasma conditions need to be sufficient to cause excitation and the medium needs to be optically thin to avoid reabsorbing the spontaneously emitted photon. Furthermore, self-absorption and line broadening influence the appearance of a line alongside intensity.[11]

Self-absorption typically occurs for emission lines in which the lower level of the transition is the ground state or close to the ground state. Emitted photons of these lines are most likely to be absorbed by neutral atoms of the same species. This often occurs at the outer, cooler layer of the plasma due to its population with atoms in the ground state. Consequently, the line is broadened and the intensity decreases. A symbolic representation of self-absorption is displayed in Figure 3. Self-absorption gets more evident with an increase in ambient gas pressure due to an increase in the local atom density.[6]

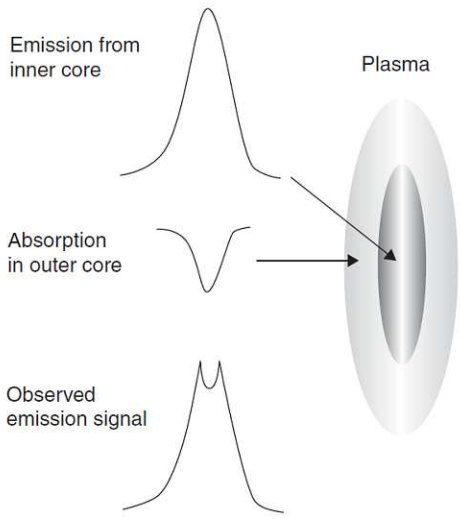


Figure 3: Schematic diagram of self-absorption[6]

Regarding line broadening, the intensity of an ideal line of a free atom is spread over a wavelength-dependent Lorentzian profile (Equation 3), where σ is the half-width.

$$I_{Lorentz}(\lambda|\lambda_0,\sigma) = A\frac{\sigma}{(\lambda-\lambda_0)^2+\sigma^2} \tag{3}$$

intensity (I); wavelength (λ); center wavelength (λ_0); standard deviation (σ); peak height (A)[12]

However, as the ideal conditions are limited to very low atomic densities, it is common for Doppler broadening to occur at the same time. Doppler broadening originates from random thermal motions of the emitting atoms and results in a Gaussian line profile (Equation 4), where $\sqrt{2ln(2)}\sigma$ is the half-width.[11] MMoreover, collisions with neutrals can lead to Lorentzian line shape broadening, while collisions with ions and electrons result in Stark broadening. Stark broadening occurs due to the split of an energy level into several sublevels in an external electric field, which is produced by the charged particles themselves. This split leads to an asymmetric distribution of the sublevels around the unperturbed level and results in asymmetric, shifted lines.[6]

$$I_{Gauss}(\lambda|\lambda_0,\sigma) = A\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{\lambda-\lambda_0}{\sigma}\right)^2} \tag{4}$$

intensity (I); wavelength (λ); center wavelength (λ_0); standard deviation (σ); peak height (A)[12]

Nevertheless, the Voigt function (Equation 5), being a convolution of a Lorentzian and Gauss profile, is commonly used to fit LIBS spectral lines.

$$I_{Voigt}(\lambda|\sigma) = \frac{1}{\pi^{\frac{3}{2}}\sqrt{2}}\int_{-\infty}^{\infty}\frac{e^{-\frac{\lambda_1^2}{2\sigma^2}}}{(\lambda-\lambda_1)+\sigma^2}d\lambda_1 \tag{5}$$

intensity (I); wavelength (λ); standard deviation (σ)[12]

However, Equation 5 has no explicit analytical formula and can only be obtained through numerical calculations. Hence, the Pseudo-Voigt profile is often used as an approximation. The Pseudo-Voigt profile (Equation 6) is a linear combination of the Lorentzian profile (Equation 3) and the Gaussian profile (Equation 4) with a weighting factor or shape coefficient.

$$I_{Pseudo-Voigt}(\lambda|\lambda_0,\sigma) = \theta\cdot A\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{\lambda-\lambda_0}{\sigma}\right)^2} + (1-\theta)\cdot A\frac{\sigma}{(\lambda-\lambda_0)^2+\sigma^2} \tag{6}$$

intensity (I); wavelength (λ); center wavelength (λ_0); standard deviation (σ), shape coefficient (θ), peak height (A)[12]

A comparison of the Lorentzian, Gaussian and Voigt profile is given in Figure 4.

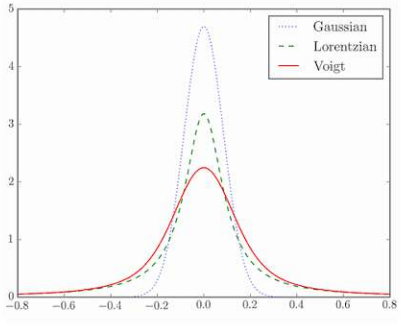


Figure 4: Comparison of Lorentzian, Gaussian and Voigt profile[13]

1.2.3 Noise

Alongside the analyte signal and continuum background, noise introduced from several sources is measured by the detector response (Equation 7).

$$I_{Detector} = I_{Signal} + I_{Background} + I_{Noise} \tag{7}$$

intensity (I)

An overview of the origins and statistical properties of noise is given in the following text.

Noise sources

In detail, noise is composed of four origins:

- Source noise:
  It is based on fluctuations in the laser-sample or laser-plasma interaction. It is characterized by a constant relative standard deviation (RSD) and affects all the spectral features.
- Shot noise:
  Fluctuations in the number of photons arriving on the detector lead to shot noise. The RSD of shot noise is proportional to the square root of the number of photons.
- Detector noise
  In LIBS, the most dominant noise source of detectors is given by photon noise or shot noise and can not be influenced by the detector itself. Other origins such as readout noise, dark current noise and photocathode dark noise are negligible.
- Instrumental (thermal) drift
  A drift is caused by a progressive laser energy increase due to the warm-up of optical and electronic components. With statistical analysis of temporal data sequences, drift effects can be recognized. It also affects the whole spectrum of emission components.[14]

Properties

Three different statistical measures can be used to describe noise:

- Stationarity - analysing changes in the variance
- Autocorrelation - focusing on serial independence of errors
- Distribution - characterising the distribution of the noise

In this thesis, the observed noise was characterised by the means of stationarity. A process is referred as stationary if attributes of a random process remain constant in a temporal context. In contrast to stationary processes, non-stationary processes do not remain constant over time or within the process and lead to a change of variance. If the variance of noise is constant, the noise is referred to as homoscedastic, otherwise, it is called heteroscedastic noise. For example, heteroscedasticity is observed in signals where noise increases with signal intensity.[15] An illustration of the behaviour of signal with homo- or heteroscedastic noise is given in Figure 5.

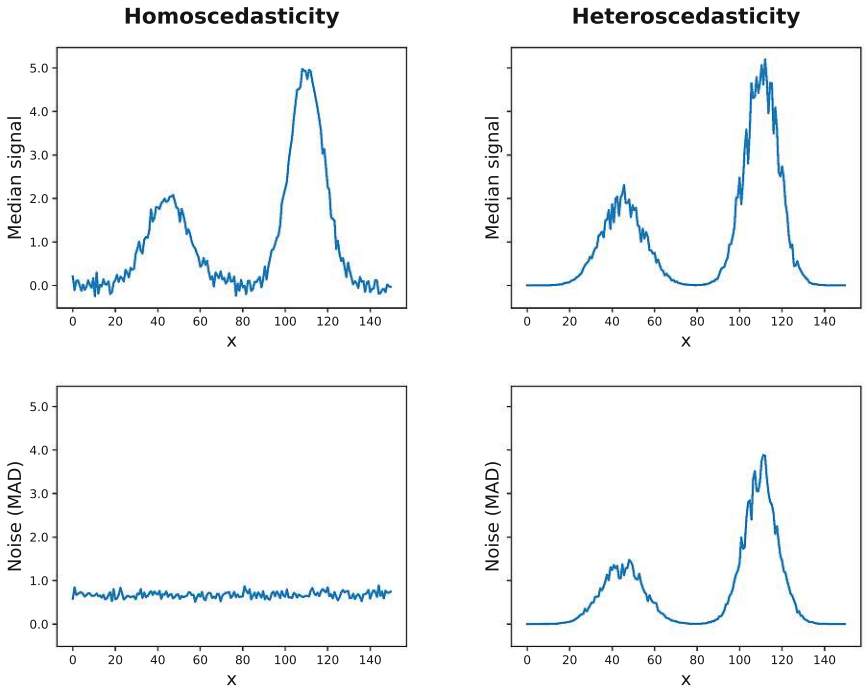


Figure 5: Comparison of homoscedastic (left) and heteroscedastic noise (right), whereas the noise was calculated as median absolute deviation (MAD) according to Equation 20

1.2.4 Measurement parameters

Albeit the model of this master thesis only depends on gate delay and laser energy, an extension of the model to other parameters is of future interest. So, the theoretical discussion of measurement parameters is not limited to the model hyperparameters. Moreover, it should be noted that isolated observations of parameter influences are not completely possible due to interactions among parameters. Subsequently, the specific experimental conditions and sample properties influence the effects of the parameters as well and a combination of factors need toe be considered when designing LIBS experiments.

Laser wavelength

The laser wavelengths influence laser-sample interactions as well as plasma-material interactions. Concerning laser-sample interactions, it was observed that shorter wavelengths lead to higher ablation rates and lower elemental fractionation. Here, the laser ablation rate is the amount of ablated mass per laser pulse per unit area and is also an indirect indicator of the coupling efficiency between the laser energy and the target material. Regarding fractionation, laser wavelength is not the most critical parameter since it is also influenced by laser energy, pulse duration and sample properties. In general, fractionation can be tuned by adjusting the laser beam irradiance.
For plasma development, the laser wavelength depends on the dominant photo absorption mechanism. Thereby, plasma ignition by inverse bremsstrahlung is favoured rather for infrared than ultra-violet (UV) wavelengths. Whereas (multi) photoionization is preferred within the UV range, longer wavelengths also facilitate plasma shielding by inverse bremsstrahlung.
Overall, infrared wavelengths are favoured for LIBS measurements because inverse bremsstrahlung also leads to reheating of the plasma which increases the lifetime and intensity of the emission lines. This surpasses the downsides of an increase in background emission and lower ablation rates as well as an increase in elemental fractionation.

Laser energy

Laser-material interaction is mainly influenced by fluence (energy per unit area,J/cm^2) and irradiance (energy per unit area and time,W/cm^2). Since those parameters include the spot size and pulse duration beside the laser energy, the quantification of laser energy alone is difficult. Yet, it can be concluded that an increase in laser energy induces higher ablation rates.
Concerning irradiance, it can be followed that the ablation rate is positively correlated with irradiance. Moreover, higher irradiance results in higher vapour density, temperatures and degree of ionizations. Emission signals also increase with rising irradiance but reach a saturation regime at some point. Overall, higher irradiance benefits analytical sensitivity.

Ambience

In general, plasma size, propagation speed, lifetime, energy and emission properties strongly depend on ambient gas and pressure.
Regarding the composition of the ambient gas, it was found that plasma lifetime increases for gases with lower conductivities, lower specific heat and higher atomic masses. This would favour argon over oxygen. Furthermore, some ambient gases can shield the sample from the laser beam if gas breakdown occurs before sample vaporization starts.
Regarding pressure, under low ambient pressure(<1 mbar), the ablated vapour can expand freely and the outer parts of the plasma get colder. Higher ambient pressures result in higher temperatures, more uniform distributions of energy within the plasma, smaller plasma sizes and consequently, a longer lifetime. Moreover, line intensity increases with increasing pressure until line broadening and self-absorption prevail.
Nevertheless, there is no universally optimal ambient gas for LIBS.

Gate delay

The delay time between the laser pulse and the measurement of the emitted light drastically influences the sensitivity of the measurement. While plasma expansion, the temperature and the intensity of the continuum emission decreases while the wavelength of the continuum emission increases. So, for short gate delays, the emission is dominated by the continuum. As continuum emission decays faster than line emission, it is reasonable to record line emission delayed to the laser pulse. However, at large delay times, the line intensities might be too low for efficient detection and the source of emissions changes from ions to atoms to molecules as a result of recombination processes at decreasing temperatures. Typical delay times reach from approximately 300 ns to >40 μs and the optimal gate delay resulting in a maximum ratio of line intensity to continuum intensity needs to be determined individually for each case.
A schematic drawing of the emission behaviour in dependency of time is shown in Figure 6.[11]

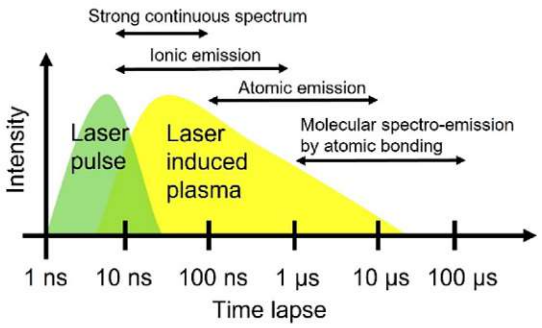


Figure 6: Schematic representation of emission processes over time[16]

1.3 Modelling

1.3.1 Multiple linear regression models

In general, a regression model aims to describe a dependent variable y with one or more independent variables $x_1, x_2, ..., x_k$. In this thesis, the spectrum intensity is the dependent variable and the laser energy (LE) and the gate delay (GD) are used as independent variables.
The general form of a regression model is given by

$$y = f(x_1, x_2, \ldots, x_k) + \varepsilon$$

or, for this thesis,

$$Intensity = f(GD, LE) + \varepsilon,$$

where f is a linear or non-linear function and ε represents the error term.

Considering linear regression models, a weighted linear combination of input variables is used as an input function with the form of

$$f(x_1, x_2, ..., x_p) = \beta_0 + \sum_{j=1}^{k} x_j \beta_j.$$

Consequently, a linear regression model is of the following shape:

$$y = \beta_0 + \sum_{j=1}^{k} x_j \beta_j + \varepsilon$$

Usually, the coefficients β_j are estimated with a given set of training data with the shape of

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \ldots & x_{1k} \\ y_2 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ y_i & x_{i1} & x_{i2} & \ldots & x_{ik} \\ \vdots & & & & \\ y_n & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} = \begin{pmatrix} y_1 & \vec{x}_1 \\ y_2 & \vec{x}_2 \\ \vdots & \\ y_i & \vec{x}_i \\ \vdots & \\ y_n & \vec{x}_n \end{pmatrix},$$

which can be separated into the target vector $\vec{y} = (y_1, \ldots, y_n)^T$ and the feature matrix $X = \begin{pmatrix} \vec{x}_1 \\ \vdots \\ \vec{x}_n \end{pmatrix}$, whereas $\vec{x}_i$ is a feature measurement for the i-th case and y_i is the corresponding observation.

So, the general aim of a regression model is to estimate each β_j as close as possible with estimated coefficients $\hat{\beta}_j$ via

$$\hat{y}_i = \hat{f}(x_{i1}, x_{i2}, \ldots, x_{ip}) = \hat{\beta}_0 + \sum_{j=1}^{p} x_j \hat{\beta}_j.$$

Polynomial regression
Polynomial regression is a specific type of linear regression model, where the independent variables are basis-expansions like $x_1 = x_1$, $x_2 = x_1^2$, ..., $x_p = x_1^p$ with an input function in the form of

$$f(x) = \beta_0 + \sum_{j=1}^{p} x^j \beta_j,$$

with polynomial degree (p). This form can also be extended to a second independent variable resulting in the model function for bilinear polynomial regression:

$$f(x_a, x_b) = \sum_{j=0}^{p} \sum_{k=1}^{p} x_a^j x_b^k \beta_{j,k} \mid j + k \leq p$$

So, the expanded form with a polynomial degree of 2 is

$$f(x_a, x_b) = \beta_{0,0} + x_a\beta_{1,0} + x_a^2\beta_{2,0} + x_b\beta_{0,1} + x_b^2\beta_{0,2} + x_a x_b \beta_{1,1}$$

and the data matrix would look like

$$\vec{y}X = \begin{pmatrix} y_2 & x_{a_{11}} & x_{a_{12}}^2 & \dots & x_{a_{1p}} x_{b_{1k}} \\ y_2 & x_{a_{21}} & x_{a_{22}}^2 & \dots & x_{a_{2p}} x_{b_{2k}} \\ \vdots & & & & \\ y_i & x_{a_{i1}} & x_{a_{i2}}^2 & \dots & x_{a_{2p}} x_{b_{2k}} \\ \vdots & & & & \\ y_n & x_{a_{n1}} & x_{a_{i3}}^2 & \dots & x_{a_{2p}} x_{b_{2k}} \end{pmatrix}.$$

1.3.2 Least-Squares Regression

For a linear regression model, the general equation can be written as

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}.$$

Here, X represents the extended feature matrix by a constant column for the intercept fit and is known as the "design matrix". The design matrix can be written in the form of

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & & & & \\ 1 & x_{i1} & x_{i2} & \dots & x_{ik} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} 1 & \vec{x}_1 \\ 1 & \vec{x}_2 \\ \vdots & \\ 1 & \vec{x}_i \\ \vdots & \\ 1 & \vec{x}_n \end{pmatrix},$$

for a target vector $\vec{y} = (y_1, \dots, y_n)^T$, a coefficient vector $\vec{\beta} = (\beta_1, \dots, \beta_n)^T$ and error terms $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$
Estimating the parameters β_j by a least-squares approach has the aim to minimize the residual sum of squares (RSS). The difference between the observed value y_i and the estimated value $\hat{y}_i$ is named as residuals. The RSS is defined in Equation 8.

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

residual sum of squares (RSS); observed value (y_i); estimated value ($\hat{y}_i$)

After minimization, the least squares estimator results in

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{y}$$

to estimate the coefficients.[17]

1.3.3 Preprocessing

In machine learning, feature transformation and feature scaling are often used as techniques to satisfy requirements of the algorithm, such as normally distributed input variables, and to boost model performance in general.[18]
Feature transformation aims to change the underlying distribution of a descriptor and feature scaling aims to provide features within the same range.

For multiple linear regression (MLR) normal distribution of variables is no prerequisite. Yet, prediction errors are supposed to follow normal distribution with a mean of 0, since the calculation of confidence intervals and variable significance is based on this assumption. Feature transformation can be used to avoid non-normally distributed errors based on skewed input variables.[19]
In terms of scale, MLR with least-squares regression does not rely on any distance relationships between descriptors like k-nearest neighbour regression. However, for polynomial regression, it can be found that algorithms for predicting the coefficients $\hat{\beta}_i$ perform better when the maximum value for each feature $x, x^2, \ldots, x^p$ is within the same orders of magnitude.[20]

So, for this thesis, transformers and scalers were implemented into the pipelines applied for regression and modelling. In the following text, the employed methods are discussed.

Yeo-Johnson Power Transformation

The Yeo-Johnson Power Transformation works with negative and positive values and is the default transformer implemented into Pythons Scikit-learn libraries "Power Transformer". The transformations are defined by Equation 9 and the reverse transformations are given by Equation 10. The power transformation parameter (ϕ) will be chosen to facilitate the best fit of a normal distribution.[21]

$$x_i^{trans}(\phi, x_i) = \begin{cases} \frac{(x_i+1)^{\phi}-1}{\phi} & if \phi \neq 0, x_i \geq 0 \\ \log(x_i+1) & if \phi = 0, x_i \geq 0 \\ -\frac{(-x_i+1)^{2-\phi}-1)}{2-\phi} & if \phi \neq 2, x_i < 0 \\ -\log(-x_i+1) & if \phi = 2, x_i < 0 \end{cases} \tag{9}$$

transformed value (x_i^{trans}); untransformed value (x_i); power transformation parameter (ϕ)

$$x_i(\phi, x_i^{trans}) = \begin{cases} (x_i^{trans} \cdot \phi + 1)^{\frac{1}{\phi}-1} & if \phi \neq 0, x_i \geq 0 \\ e^{x_i^{trans}} - 1 & if \phi = 0, x_i \geq 0 \\ 1 - (-(2-\phi))^{x_i^{trans}} + 1)^{\frac{1}{2-\phi}} & if \phi \neq 2, x_i < 0 \\ 1 - e^{-x_i^{trans}} & if \phi = 2, x_i < 0 \end{cases} \tag{10}$$

transformed value (x_i^{trans}); untransformed value (x_i); power transformation parameter (ϕ)

Standard Scaler

With this scaler, the values are scaled in a way that their mean equals 0 and the variance is 1. The formula of the standard scaler is given by Equation 11.[22]

$$x_i^{scaled} = \frac{x_i - \overline{x}}{s_d(x)} \tag{11}$$

scaled value (x_i^{scaled}); untransformed value (x_i); mean value ($\overline{x}$); standard deviation (s_d)

1.3.4 Model evaluation

In order to analyse the quality of a model, several performance metrics and visualizations for residual assessment are available. The most important ones for this thesis will be discussed subsequently.

Coefficient of determination

The most prominent metric is the coefficient of determination (R^2), which represents how much the error of prediction can be reduced, in per cent, by using the model predictions instead of the mean value of the targets. So, for calculation, total sum of squares (TSS) (see Equation 12) and RSS (see Equation 8) are connected in Equation 13 for R^2.

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y}_i)^2 \tag{12}$$

total sum of squares (TSS); observed value (y_i); mean value ($\overline{y}_i$)

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{13}$$

coefficient of determination (R^2); residual sum of squares (RSS); total sum of squares (TSS)

R^2 usually lies between 0 and 1, whereas 1 is the optimum. Despite the popularity of this measure, a major drawback of it is, that it rises by simply increasing the number of predictors even if the new variable does not contribute to the model at all. To address this issue other metrics or the adjusted R^2 can be used.
Moreover, R^2 can be diminished to the square of the Pearson correlation coefficient (PCC) if only one independent variable is present.[15]

Spearman's Rank Coefficient

The Spearman's rank correlation coefficient (ρ_S) is a metric to measure the monotonic relationship between two, paired variables and a robust alternative to the PCC in terms of outliers. Throughout calculation, each variable is converted into ranks and the rank difference (d_r) for each variable pair is calculated and subjected to Equation 14.

$$\rho_S = 1 - \frac{6 \sum_{i=0}^{n} d_{r_i}^2}{n(n^2 - 1)} \tag{14}$$

Spearman's rank correlation coefficient (ρ_S); rank difference (d_r); number of observations (n)

Like the PCC, it can take values between -1 to 1, where 0 indicates no association between the variables.[23]

F-value

Compared to R^2, a better measure of fit is the F-value or F-ratio. The F-ratio is used in hypothesis testing to confirm that at least one coefficient of $\beta_0, \beta_1, ..., \beta_k$ is significantly different from 0. In detail, the null hypothesis and alternative hypothesis are in the form of

$$H_0 : \beta_0, \ldots, \beta_k = 0$$
$$H_A : \exists\ \beta \in \{\beta_0, \ldots, \beta_k\} : \beta \neq 0$$

As a test statistic, the F-value (Equation 17) is used which is the ratio of the regression mean square (MSR) (Equation 15) and mean square error (MSE) (Equation 16). However, the F-value can also be calculated with R^2.

$$\text{MSR} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{k} \tag{15}$$

regression mean square (MSR); estimated value ($\hat{y}_i$), mean value ($\overline{y}_i$); number of independent variables (k)

$$\text{MSE} = \frac{\text{RSS}}{n-k-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1} \tag{16}$$

mean square error (MSE); residual sum of squares (RSS); observed value (y_i); mean value ($\overline{y}_i$); number of observations (n)

$$\text{F-value} = \frac{\text{MSR}}{\text{MSE}} = \frac{R^2(n-k-1)}{(1-R^2)k} \tag{17}$$

regression mean square (MSR) (Equation 15); mean square error (MSE) (Equation 16); coefficient of determination (R^2)

For rejecting the null hypothesis, the probability is received by comparing the F-ratio to an F distribution. Nevertheless, the F-ratio itself can be used as a measure for the fit of the model. The higher the value of the F-ratio, the better fits the model to the data.[15,24]

Analysis of residuals

Visual examination of results often provides a more insightful understanding of the data compared to numerical residual analysis. By graphical representation, outliers, misfits and scedasticity are well interpretable. For this thesis, histograms of residuals, as well as residual versus number of observations plots, were used.

1.3.5 Feature Selection

To optimize models with more than one descriptor, feature selection is a necessary step. In this thesis, feature selection was included in the modelling pipeline via stepwise regression with subsequent elimination of non-significant coefficients. If no model fit with significant coefficients was possible, an empty model was returned. A general variable selection workflow is depicted in Figure 7.

Stepwise regression

Stepwise regression can be described as a loop of forward selection with subsequent backward selection until no better set of variables could be found. As scoring criteria, the F-value of the model was used in this thesis. A brief overview of the algorithm follows.[15]

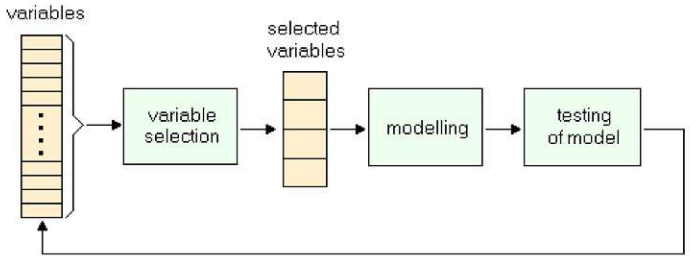


Figure 7: Schematic drawing of a feature selection process[15]

Forward selection:

- At initialisation: select the variable which results in the best model score
- Test all combinations with the remaining variables to find a combination which outreaches the current best model score
- If the model score could be improved, update the best variable selection.

Backward selection:

- Test all combinations omitting one of the preselected variables
- If the model score could be improved, update the best variable selection.

If a previous combination is again identified, exit the stepwise regression algorithm. Otherwise, continue with forward selection.

If only hierarchical polynomials (including all monomials up to the highest degree) should be modelled, the backward step is omitted and forward selection was limited to testing coefficients in ascending order of degrees.

Selection of significant coefficients

To test if a coefficient β_j is significantly different from zero, a t-Test is performed. The null and alternative hypotheses can be written as

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

The t-statistic of a coefficient is determined by Equation 18 and is then compared to the t-distribution to achieve the p-value. If the p-value is above the significance level, the null hypothesis is rejected.[24,25]

$$\text{t-statistic} = \frac{\hat{\beta}_j}{s_e(\beta_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{1}{n-2} \cdot \frac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \tag{18}$$

predicted coefficient ($\hat{\beta}_j$); standard error (s_e); mean square error (MSE) (Equation 16);number of observations (n); value of independent variable (x_i); mean of independent variable x ($\overline{x}$)

For feature selection, the following algorithm was performed:

- The p-values for all coefficients are calculated.
- If the highest p-value is above the significance level of 0.05, the coefficient is omitted. Otherwise, feature selection is finished and the algorithm is exited.
- The model is refitted with the remaining coefficients and the algorithm is repeated.

In the special case of hierarchical polynomials, the algorithm is exited only if the coefficient with the highest degree is significantly different from zero. The other coefficients are tolerated to be non-significant to sustain the structure of hierarchical polynomials.

2 Datasets

Datasets of three different materials (low-alloyed steel, aluminium alloy and borosilicate glass) were used as a foundation to build prediction models for these materials. The datasets were provided by David Prochazka et al. from the Brno University of Technology and were measured on a Lightigo FireFly LIBS. The experimental settings are given in Table 1. An overview of the provided data is given in Table 2. The measured materials were certified reference materials with defined sample compositions (see Table 3).[26]

Table 1: Applied experimental settings for the measurement of the supplied datasets on a Lightigo FireFly LIBS[26]

Parameter	Settings
Laser	Nd:YAG (20 Hz, 532 nm, 10 ns)
Spot size	100 µm diameter
Optical cable	400 µm diameter
Spectrometer	Echelle-type
Detector	EMCCD

Table 2: Overview of the structure of the provided datasets

	low alloyed steel	aluminium alloy	borosilicate glass
Sample name	SUS-1R	ERM-EB316	NIST-1411
No. of conditions	15	48	64
Replicates/condition	100	100	100
Range of laser energy (mJ)	5-75	5-90	20-140
Range of gate delay (ns)	200-4000	100-20000	500-2600

Table 3: Composition of provided samples

	mass fraction (%)		
Element	SUS-1R	ERM-EB316	NIST-1411
Al	-	87.47	-
Al_2O_3	-	-	5.68
B_2O_3	-	-	10.94
BaO	-	-	5.00
C	0.9	-	-
CaO	-	-	2.18
Co	0.3	-	-
Cr	1.7	-	-
Cu	0.7	0.0297	-
Fe	88.9	0.1054	-
Fe_2O_3	-	-	0.05
K_2O	-	-	2.97
Mg	-	0.045	-
MgO	-	-	0.33
Mn	1.1	0.204	-
Mo	0.9	-	-
Na_2O	-	-	10.14
Nb	0.55	-	-
Ni	2.9	0.0235	-
P	0.02	-	-
S	0.017	-	-
Si	0.8	11.98	-
SiO_2	-	-	58.04
SrO	-	-	0.09
Ti	-	0.079	-
TiO_2	-	-	0.02
V	0.5	-	-
W	0.7	-	-
Zn	-	0.0611	-
ZnO	-	-	3.85

3 Development

In the following, the applied principles will be discussed with the dataset of low-alloyed steel (SUS-1R) first. For the aluminium alloy (ERM-EB316) and borosilicate glass (NIST-1411) dataset, only deviations will be discussed. Moreover, measured data will be displayed as blue and simulated data as red unless otherwise stated.

3.1 General

3.1.1 Modelling considerations

According to Equation 7, the detector response is the summation of signal, continuum background and noise. Considering, a typical LIBS spectrum, the continuum background is represented as a baseline. Thereby, Equation 7 can be rewritten as Equation 19. Consequently, the main idea of the development process was to model noise, baseline and signal separately. A schematic representation of this idea is shown in Figure 8. In the following, the modelling of these three parts will as well as the final merge into one model be discussed intensively.

$$I_{Detector}(\lambda) = I_{Signal}(\lambda) + I_{Baseline}(\lambda) + I_{Noise}(\lambda) \tag{19}$$

intensity (I); wavelength (λ)

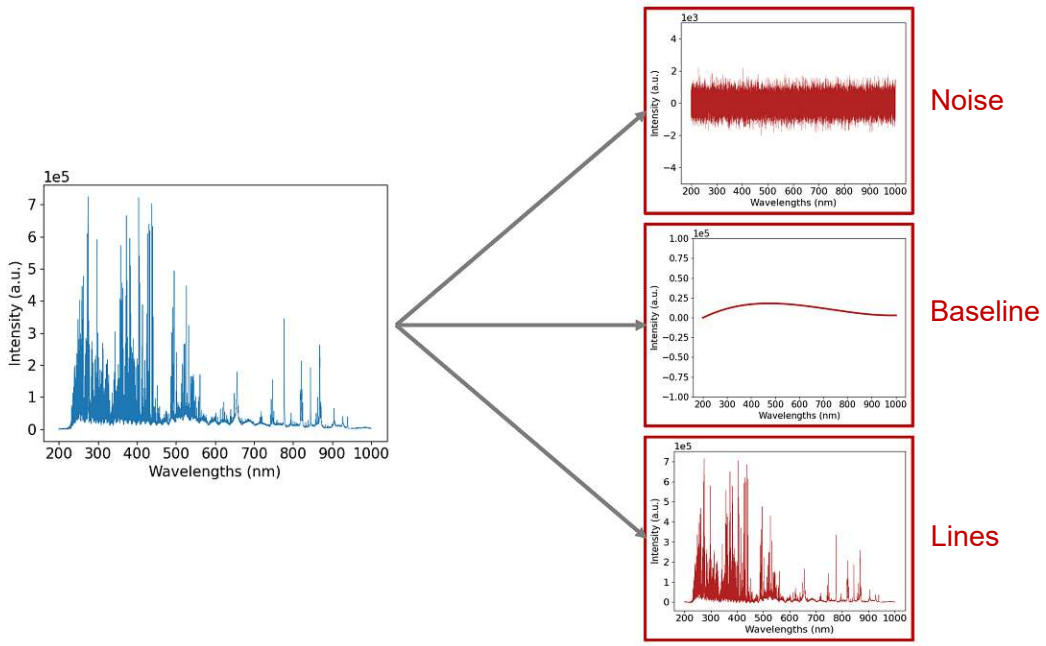


Figure 8: Schematic drawing of a separation of a measured spectrum into noise, baseline and signal according to Equation 19; Plots: I (a.u.) vs. λ (nm)

3.1.2 Code

The entire development process was calculated in Python 3.11.1. In detail, the following published packages were used:

- `numpy` 1.24.3[27]
- `pandas` 2.0.1[28,29]
- `scikit-learn` 1.2.2[30]
- `statsmodels` 0.14.0[31]
- `lmfit` 1.2.1[32]
- `matplotlib` 3.7.1[33]
- `seaborn` 0.12.2[34]

Moreover, several custom modules were designed in compliance with object-oriented programming principles as a supplement:

- `datalayers.py`
  This is the main module of the thesis. Here, a dataset is separated into three classes (`Dataset`, `Condition`, `Spectrum`) representing the information layers which are connected by composition. Within this module, performing steps such as preprocessing, regression and modelling is possible for all three classes.
- `model.py`
  With this module, the models for baseline coefficients and elemental lines as well as the overall baseline and the complete spectrum can be calculated.
- `transformer.py`
  This module includes custom transformers for shifting data, log-scale transformations and bidirectional feature selection.
- `regressor.py`
  Here, a wrapper of `statmodels.py`'s weighted least squares regression, to make this estimator compatible with the `scikit-learn` framework, is included.
- `combalg.py`
  The functions of this module were provided by Zuzana Gajarska and were used to identify spectral lines.
- `utils.py`
  This module includes several utility functions like the Gaussian, Lorentzian or Pseudo-Voigt distribution functions.

3.2 Preprocessing of raw data

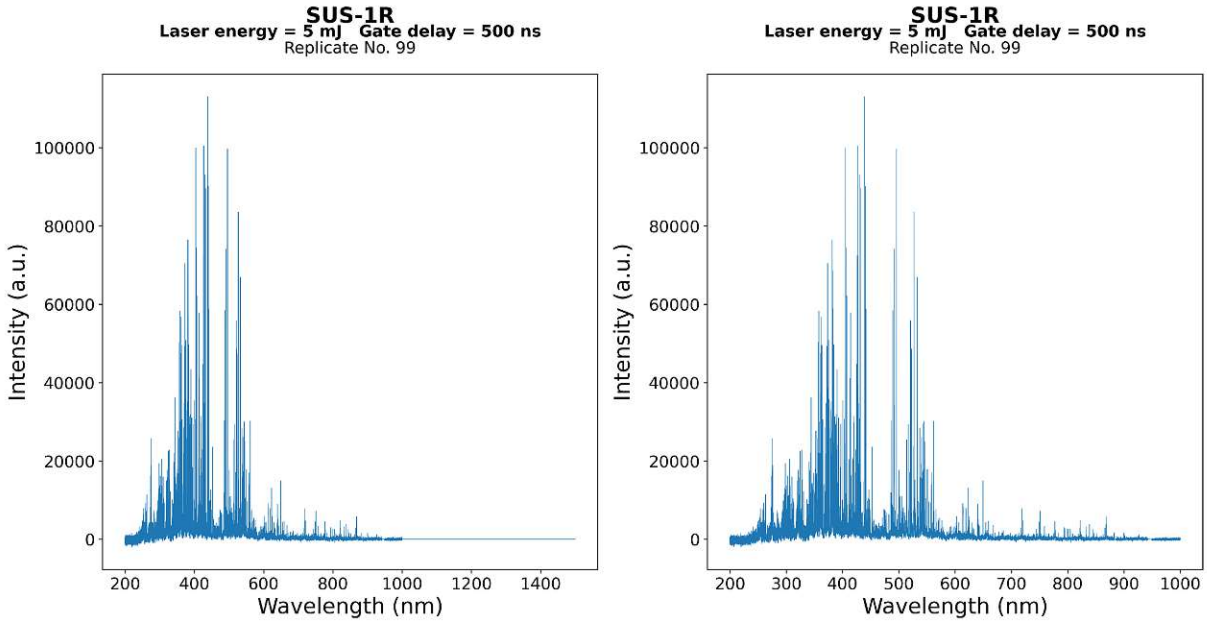
Before introducing data to modelling, preprocessing to convert the provided text files and to remove faulty signals had to be performed.

3.2.1 Initialization of dataset

The provided datasets were comprised of separate text files, with each file containing measurements from all 100 replicate measurements for a particular condition.. To enhance computation speed, the files of each dataset were merged and stored as a `Dataset` pickle object initially and reloaded when used.

3.2.2 Interpolation and removal of faulty wavelengths

In certain wavelength ranges across all spectra in the dataset, intensity measurements resulted in zero values. To facilitate further data processing, the affected range was linearly interpolated using the two values preceding and following the range.
This was implemented as `Condition(...).interpolate_faulty_wavelengths(...)`. To avoid interpolation in the case of zero being the real measured intensity, it was checked that zero intensity occurs over all replicate measurements for this wavelength. This was implemented into the stated method as well. Moreover, registered wavelengths with no measured intensity at the end of the spectrum were deleted as well.
As an example, the effect of preprocessing can be seen in Figure 9 for a low-alloyed steel sample. Here, interpolation took place from 883.08-885.48 nm and 941.94-948.76 nm and redundant pixels were removed at the end of the spectrum.



(a) before preprocessing **(b)** after preprocessing

Figure 9: LIBS spectrum of low-alloyed steel (SUS-1R) recorded at 5 mJ laser energy and 500 ns gate delay (replicate 99) before (a) and after preprocessing (b). Interpolation took place from 883.08-885.48 nm and 941.94-948.76 nm and redundant pixels were removed at the end of the spectrum.

3.3 Noise

3.3.1 Characterisation

Before modelling, the present noise of the spectra had to be analysed concerning its properties. As discussed in section 1.2.3, stationarity was the focus of the characterisation. For stationarity assessment, the noise was calculated for all conditions of a dataset and plots according to Figure 5 were used to evaluate the scedasticity.
As noise, the MAD used. Therefore, the median intensity of a measurement condition ($\tilde{I}_c$) over all replicate measurements was calculated by

$$\tilde{I}_c(\lambda) = \text{median}_{i=0}^{99}\left(I_i(\lambda)\right)$$

for each wavelength at first. Afterwards, the MAD is calculated by the median of the difference of $\tilde{I}_c$ and every single measurement according to Equation 20.

$$\text{noise}(\lambda) = \text{median}_{i=0}^{99}\left(I_i(\lambda) - \tilde{I}_c(\lambda)\right) \tag{20}$$

wavelength (λ); intensity (I); median intensity of a measurement condition ($\tilde{I}_c$)

3.3.1.1 Low-alloyed steel (SUS-1R)

Comparing the noise vs. wavelength plots in Figure 10 to Figure 5, the noise behaviour of the low-alloyed steel sample clearly shows heteroscedasticity. For conditions with overall low signal intensity, heteroscedastic characteristics are blended in with homoscedastic characteristics due to the fact that less signal is present.

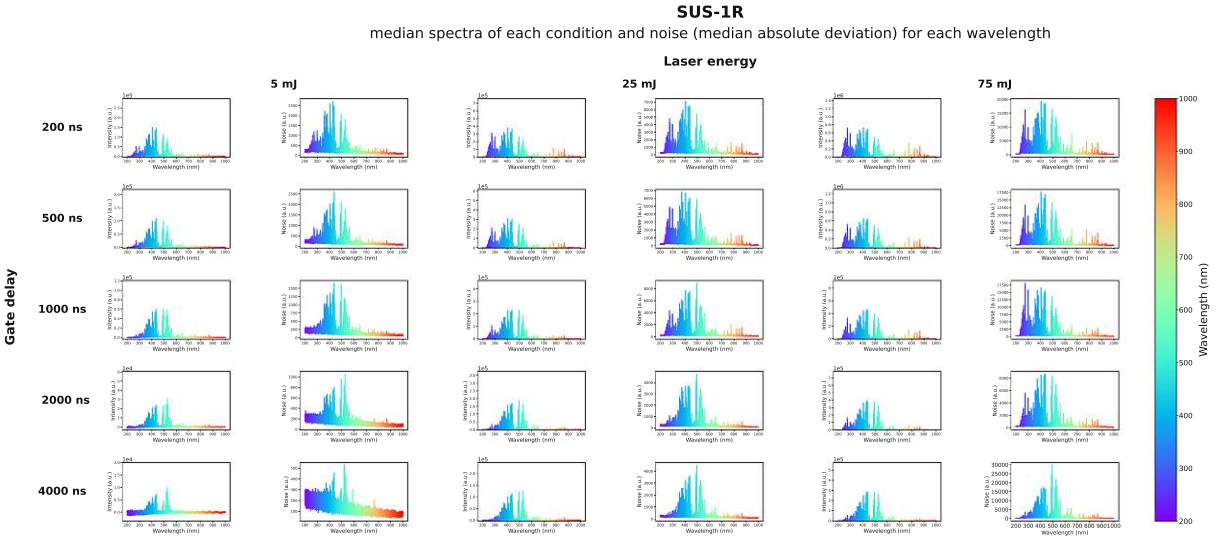


Figure 10: Low-alloyed steel (SUS-1R): median LIBS spectra ($\tilde{I}_c$ (a.u.) vs. λ (nm)) and noise spectra (MAD (a.u.) vs. λ (nm)) of each condition

3.3.1.2 Aluminium alloy (ERM-EB316)

Just as with low-alloyed steel, heteroscedasticity of noise is visible in Figure 11.

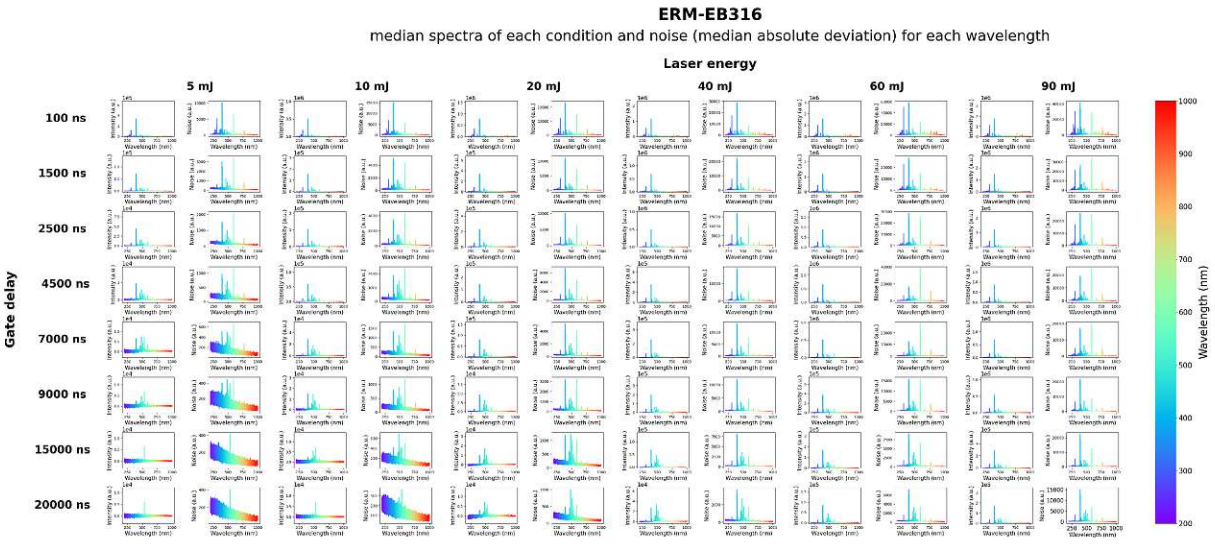


Figure 11: Aluminium alloy (ERM-EB316): median LIBS spectra ($\tilde{I}_c$ (a.u.) vs. λ (nm)) and noise spectra (MAD (a.u.) vs. λ (nm)) of each condition

3.3.1.3 Borosilicate glass (NIST-1411)

Same as for low-alloyed steel, heteroscedasticity of noise can be seen in Figure 12.

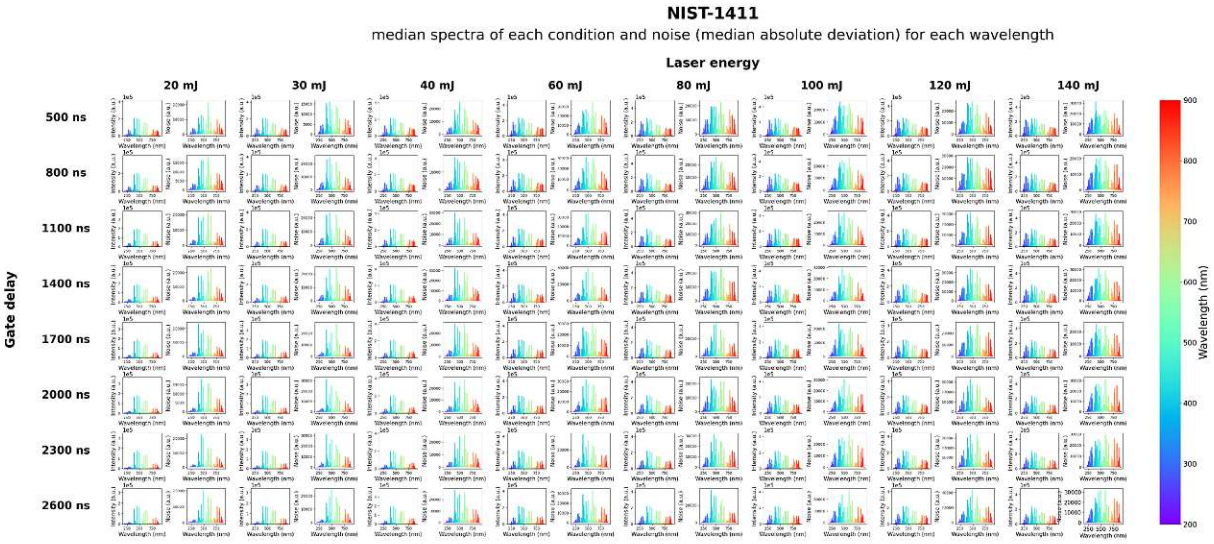
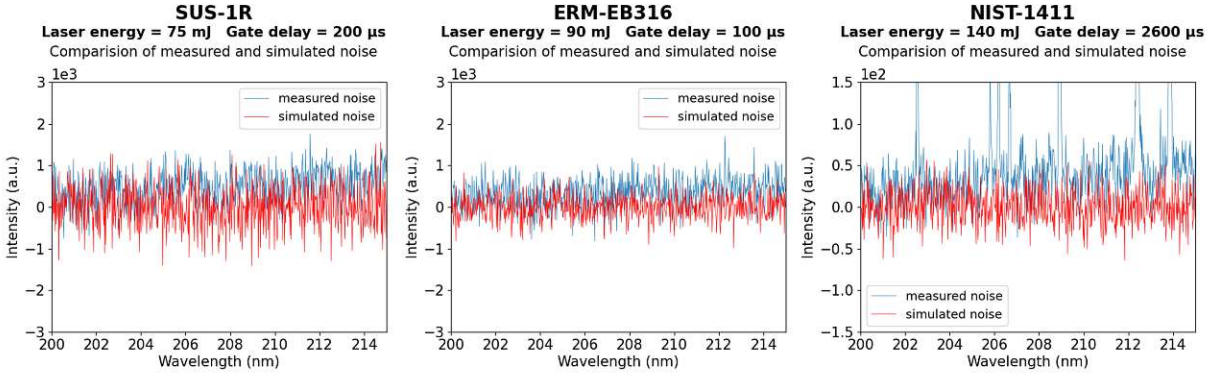


Figure 12: Aluminium alloy (ERM-EB316): median LIBS spectra ($\tilde{I}_c$ (a.u.) vs. λ (nm)) and noise spectra (MAD (a.u.) vs. λ (nm)) of each condition

3.3.2 Modelling

Even though the present noise was identified as heteroscedastic, the model was cut down to normally distributed random numbers since the correctness of simulated noise only has a minor impact onto the overall model quality. As a model parameter, it was opted for an absolute intensity value which will be used as the standard deviation of the Gaussian distribution. In Figure 13 you can see that the simulated noise fits to measured data.



(a) Low-alloyed steel (SUS-1R)
λ_{range}: 200-215 nm;
LE: 75 mJ; GD: 200 ns

(b) Aluminium alloy (ERM-EB316)
λ_{range}: 200-215 nm;
LE: 90 mJ; GD: 100 ns

(c) Borosilicate glass (NIST-1411)
λ_{range}: 200-215 nm;
LE: 140 mJ; GD: 2600 ns

Figure 13: Comparison of measured (blue) and simulated noise (red) for low-alloyed steel (a), aluminium alloy (b) and borosilicate glass (c); Plots: I (a.u.) vs. λ (nm)

3.3.3 Chapter summary

For all materials, heteroscedastic noise is present. However, the modelling effort was cut down to normally distributed random numbers which equals the assumption of homoscedastic noise. This was performed by taking into consideration that the noise simulation quality does not impact the overall model quality significantly. Moreover, it was found that the simulated noise fits well with the measured data for all three materials.

3.4 Baseline

The goal was to build a bilinear baseline model in dependency on gate delay and laser energy. This baseline model is supposed to be a combination of bilinear models in dependency on gate delay and laser energy of the coefficients of the underlying hierarchical baseline polynomials (see Equation 22 and 23). The underlying baseline polynomials are calculated by baseline regression of measured data.
In particular, it was opted for polynomials due to their simple calculation and well-manageable coefficients. Moreover, hierarchical polynomials have the benefit to build cross-terms while modelling.[35] Splines and the Liebers algorithm were taken into consideration as well. However, the reproducible calculation of spline coefficients is difficult and the Liebers algorithm only returns a numeric list of values.
In the following, the steps of regression and modelling are intensively discussed.

3.4.1 Regression

Identifying pivot points

Prior to baseline regression, anchor points of the baseline had to be identified. Therefore, a custom moving minimum filter was applied as part of the method `Spectrum.Baseline.find_pivot_points(...)`. Here, the window width of the minima filter (WW_{min}) had to be fine-tuned. The step width was set to half the window width. The general working principle of the minima filter is shown in Figure 14.

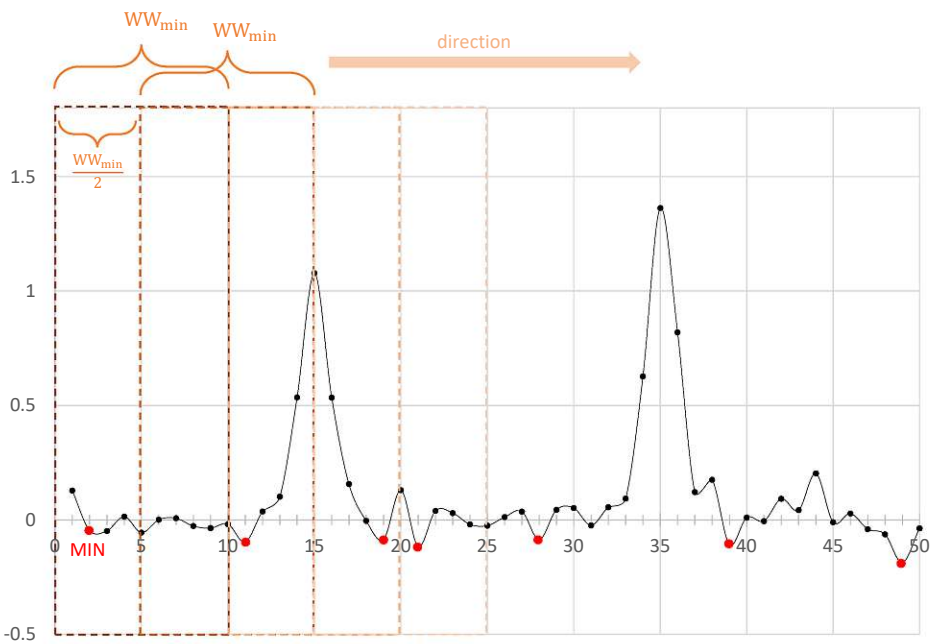


Figure 14: Schematic representation of a minima filter with WW_{min}=10 and a step width of $\frac{WW_{min}}{2}$=5 including detected minima (red)

While optimising pivot point identification, a few more parameters were introduced:

- Detailed wavelength range (DWR):
  To avoid under-fitting at the start end of the spectrum, the step width of the filter was decreased within a specific wavelength range at the upper and lower ends of the spectrum. The shrinkage was determined by a parameter named detail factor (DF). Thereby, the step width can be changed from $\frac{\mathrm{WW_{min}\ (nm)}}{2}$ to $\frac{\mathrm{WW_{min}\ (nm)}}{2 \cdot \mathrm{DF}}$. An example of an implementation of a minima filter with a DWR is given in Figure 15a. Comparing Figure 15a and 14, an increase of detected minima at the start and end of the spectra can be seen.
- Detail factor (DF):
  As discussed above, the detail factor determines the step width of the filter within the detailed wavelength region at the boundaries of the spectrum
- Reverse direction (REVD):
  It was found that the amount of pivot points at both spectral ends is not always well balanced. So, the option to run the minimum filter in both directions was developed to improve the fit if needed.

Moreover, the intensity value of the minimum points was set to the median of four values before and four values after the minimum point to provide a more realistic intercept of the polynomial baseline (see Figure 15b).

Regressor

It was found that simple least-squares regression does not yield satisfying fits for all spectra. For some, weighted least-squares regression with increased weight for the first and last three data points was necessary to avoid misfits at the start or end of the spectrum. So, the endpoint-weight (EPW) was introduced as an additional tuning parameter.

Feature selection

To select the best polynomial degree, a custom feature selector was built according to section 1.3.5 and is included in the transformer.py module. The selector was also designed to be compatible with the scikit-learn framework. For baseline regression, the polynomial degree was limited between three and five to avoid over- or underfitting and round off errors. For significance selection, the level of significance was set to 0.05.

Shifted polynomials

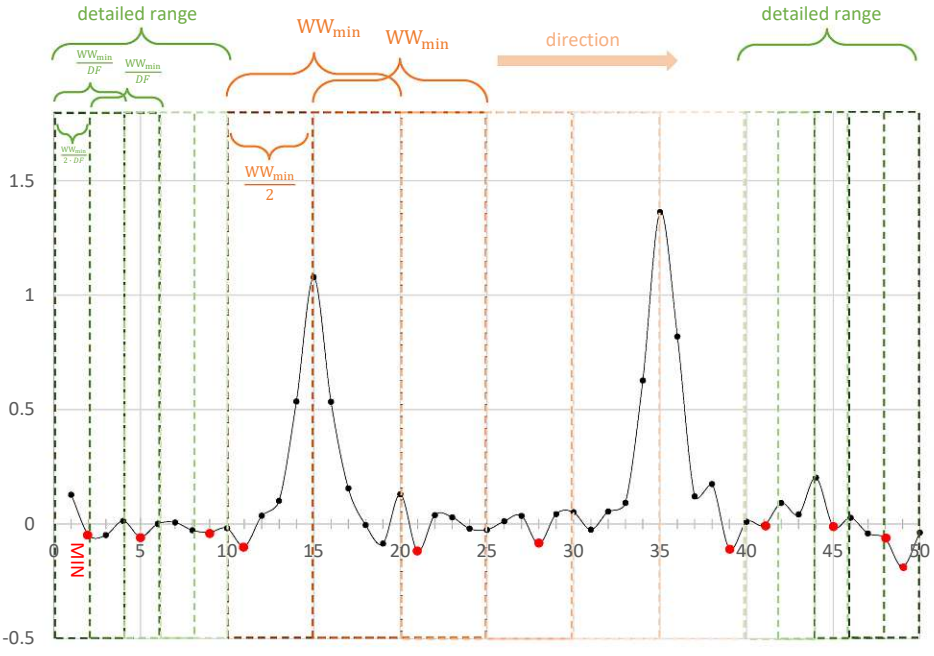
A problem faced during development was that polynomials which fitted quite well visually achieved high p-values for their coefficients. This circumstance was caused due to a numerical problem when calculating high potences of the wavelengths as independent variables. Values to a higher power lead to big numerical values and big values show higher collinearity between them, even though the same calculations with lower x-values would not. To avoid such false results, the independent variables, here the wavelengths, were centred around their mean. To do so, a custom transformer was included in the transformer.py module. So, the final baseline polynomial is in the shape of Equation 21.

$$I_{baseline}(\lambda) = a_0 + a_1 \cdot (\lambda - \overline{\lambda}) + a_2 \cdot (\lambda - \overline{\lambda}) + \ldots + a_p \cdot (\lambda - \overline{\lambda})^p \quad 3 \leq p \leq 5 \tag{21}$$

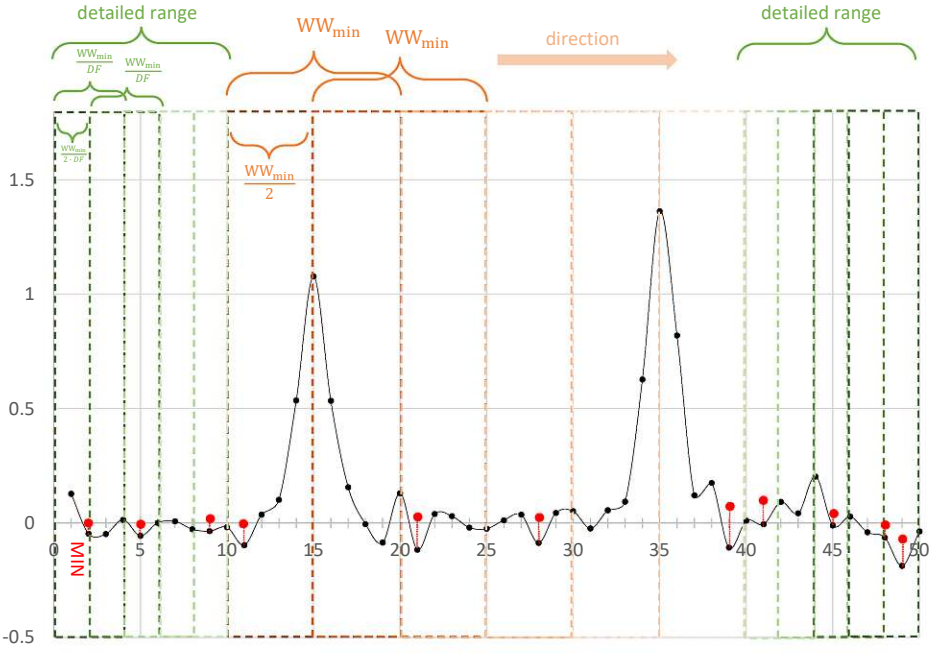
intensity (I); wavelength (λ); mean value of the wavelength range ($\overline{\lambda}$); coefficient (a); polynomial degree (p)

Optimisation

Overall, several parameters had to be optimised to provide a confident baseline fit for all materials. The starting point for optimisation were the following default settings:
$\mathrm{WW_{min}} = 40$ nm; DWR = 0 nm; DF = 1; REVD = false; EPW= 1
The parameter optimisation process was performed with a single spectrum of a representative measurement condition for the dataset. Yet, the parameters needed to be confirmed for the entire dataset. This is discussed in the first modelling step (section 3.4.2). Subsequently, the parameter optimisation for each material is discussed.



(a) without median signal values

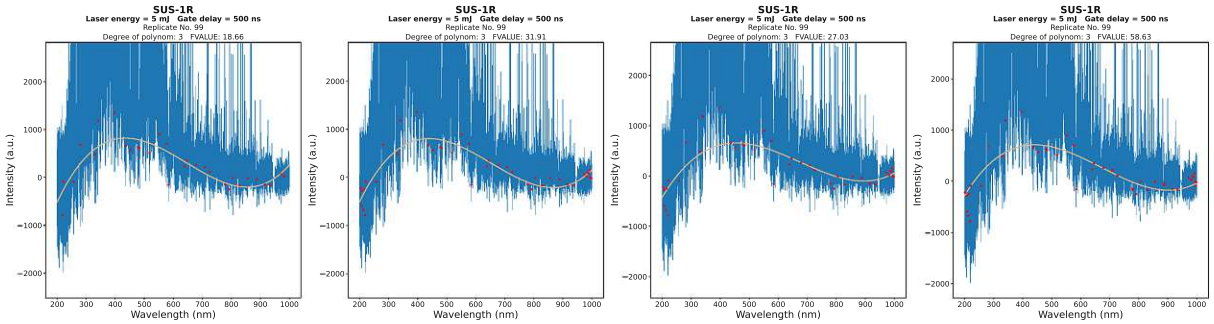


(b) with median signal

Figure 15: Schematic representation of a minima filter with WW_{min}=10 and a step width of $\frac{WW_{min}}{2}$=5 in the regular region (orange) and a DWR of 10 at the start and end (green) with a DF of 2.5 including detected minima (red); Figure 15a shows the detected minima at their given signal, 15b shows the signal values of the minima set at the median value of four values before and after the minimum point

3.4.1.1 Low-alloyed steel (SUS-1R)

Using only default options in Figure 16a resulted in severe underfitting at the upper end of the spectrum due to a lack of appropriate pivot points. So, increasing the pivot points by changing the settings of DWR to 20 nm and DF to 10 improved this issue slightly (Figure 16b). Next, the EPW was set to 10000 to force the polynomial through the endpoints (Figure 16c). This helped the fit for the given example. However, due to an imbalanced distribution of pivot points at the start and end of the spectrum, as seen in Figure 17, the REVD was set to true to increase the number of pivot points at the upper end (Figure 16d). This improved the fit at the upper end in Figure 16d compared to Figure 16c significantly. The determined degree of the polynomial remained three independent of the parameters.



(a) WW_{min}: 40 nm; DWR: 0 nm; DF: 1; REVD: false; EPW: 1; polynomial degree: 3

(b) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 1; polynomial degree: 3

(c) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 10000; polynomial degree: 3

(d) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: true; EPW: 10000; polynomial degree: 3

Figure 16: Low-alloyed steel (SUS-1R): optimisation steps for baseline fit parameters starting with default settings (a), intermediate settings (b, c) and ending with final settings (d) demonstrated for 5 mJ laser energy and 500 ns gate delay (replicate 99); Plots: I (a.u.) vs. λ (nm) with measured data (blue), calculated baseline (beige) and pivot points (red)

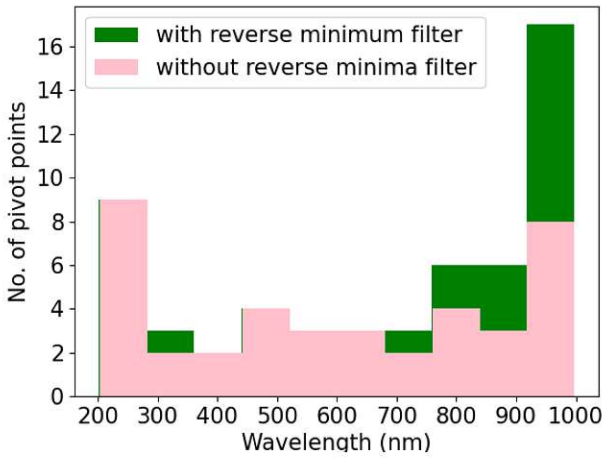
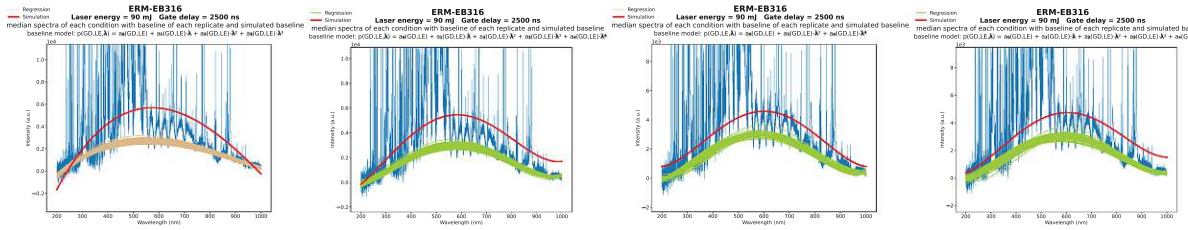


Figure 17: Low-alloyed steel (SUS-1R): distribution of pivot points with (green) and without (pink) bidirectional application of the minimum filter (reffering to Figure 16c and 16d)

3.4.1.2 Aluminium alloy (ERM-EB316)

For aluminium alloy, it was not possible to determine the best baseline settings by visual analysis of the baseline, since it did not change significantly with the change of a parameter. However, at baseline modelling (see section 3.4.2), it was observed that the influence of baseline regression settings had a strong impact on the simulated baseline. So, the simulated baseline was taken into account to find the optimal settings for baseline regression. Since baseline modelling required a predetermined degree of polynomial, a degree of four was chosen because a degree of three led to severe under-fitting, as seen in Figure 18a. For default conditions, no baseline fit was possible. That's why the settings for DWR and DF were taken over from the low-alloyed steel sample (Figure 18b). As for the steel sample, an increase of EPW to 10000 improved the fit significantly (Figure 18c). However, the attempt to apply the minimum filter in both directions did not result in improvements (Figure 18d). So, the best baseline regression settings were defined according to Figure 18c. The offset of the calculated baseline is accepted since it is constant over the wavelength range.



(a) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 1; polynomial degree: 3

(b) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 1; polynomial degree: 4

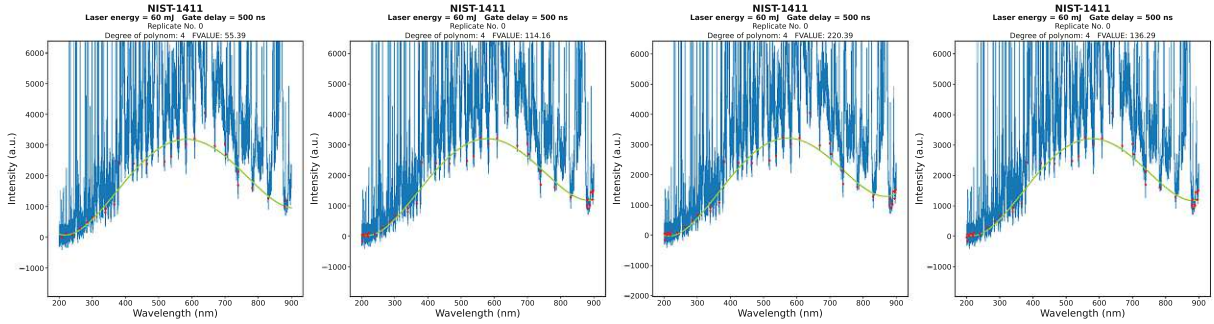
(c) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 10000; polynomial degree: 4

(d) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: true; EPW: 10000; polynomial degree: 4

Figure 18: Aluminium alloy (ERM-EB316): optimisation steps for baseline fit parameters starting with default settings (a, b), final settings (c) and intermediate settings (d) demonstrated for 90 mJ laser energy and 2500 ns gate delay including simulated baseline (red) and regressed baseline (beige/green) and median spectra (blue) for each replicate; Plots: I (a.u.) vs. λ (nm)

3.4.1.3 Borosilicate glass (NIST-1411)

Compared to default settings (Figure 19a), changing DWR to 20 nm and DF to 10 improved the fit by an increase of pivot points (Figure 19b). An adjustment of EPW (Figure 19c) and REVD (Figure 19d) did not have an impact on the baseline. So, it was settled with the settings of Figure 19b. Additionally, the polynomial degree of four did not change throughout optimisation.



(a) WW_{min}: 40 nm; DWR: 0 nm; DF: 1; REVD: false; EPW: 1; polynomial degree: 4

(b) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 1; polynomial degree: 4

(c) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: false; EPW: 1000; polynomial degree: 4

(d) WW_{min}: 40 nm; DWR: 20 nm; DF: 10; REVD: true; EPW: 1; polynomial degree: 4

Figure 19: Borosilicate glass (NIST-1411): optimisation steps for baseline fit parameters starting with default settings (a), final settings (b) and intermediate settings (c, d) demonstrated for 60 mJ laser energy and 500 ns gate delay (replicate 0); Plots: I (a.u.) vs. λ (nm) with measured data (blue), calculated baseline (green) and pivot points (red)

3.4.2 Modelling

The aim of this section was to build a bilinear baseline model based on gate delay and laser energy. So, Equation 21 is transformed to Equation 22 by including models for each coefficient in dependency of gate delay and laser energy. For coefficient models, the maximum degree was limited to two, as seen in Equation 23. Prior to discussing the models of each material, the general workflow is described below.

$$I_{baseline\ model}(GD, LE, \lambda) = \sum_{i=0}^{p} a_i(GD, LE) \cdot (\lambda - \overline{\lambda})^i \quad 3 \leq p \leq 5 \tag{22}$$

intensity (I); laser energy (LE); gate delay (GD); wavelength (λ); mean value of the wavelength range ($\overline{\lambda}$); coefficient (a); polynomial degree (p)

$$a_i(GD, LE) = a_{im0} + a_{im1} \cdot GD + a_{im2} \cdot GD^2 + a_{im3} \cdot GD \cdot LE + a_{im4} + a_{im5} \cdot LE + \cdot LE^2 \tag{23}$$

coefficient (a); gate delay (GD); laser energy (LE)

Common polynomial degree

To build the baseline model, it was necessary to get a baseline polynomial of each spectrum of the dataset including the same coefficients in advance. Therefore, the distribution of polynomial degrees for a dataset was analysed and the dataset was refitted with the median degree. For analysis, tools such as pie plots were used. To visualise data of an entire dataset, the method `Dataset.plot_dataset(...)` was applied. However, it was not possible to fit a baseline with the set common degree to each spectrum. Spectra with no fit were omitted prior to further processing.

Preprocessing

Median coefficients
Omitting spectra due to unsuccessful baseline fit resulted in an imbalance of available baseline information for each condition. So, the median for each coefficient of each condition was used as targets in the feature matrix.

Transformation and Scaling
The feature matrix and targets were transformed and normalized. For feature columns (intercept, LE, GD, LE^2, GD^2, $LE \cdot GD$), the intercept column, made up of ones, was not transformed and scaled. Moreover, the Yeo-Johnson transformation (Equation 9) with subsequent zero-mean, unit-variance normalisation was applied as part of scikit-learns `PowerTransformer`.
However, the application of the `PowerTransformer` onto the target column (the values of each coefficient a_i) was not possible, since inverse transformation resulted in a division by zero due to rounding errors for some targets. Therefore, a custom log-transformer was built in the form of Equation 24. Within this custom transformer, the median sign of the values was calculated (Equation 25) and used for the inverse transformation according to Equation 26. Furthermore, this custom transformer was combined with scikit-learns `StandardScaler`.

$$y_i^{trans}(y_i) = \log(|y_i|) \tag{24}$$

transformed value (y_i^{trans}); untransformed value (y_i)

$$\tilde{\delta}(y) = \text{median}_{i=0}^{n}(\text{sgn}(y_i)) \tag{25}$$

median sign ($\tilde{\delta}$); sign function (sgn)

$$y_i\left(\tilde{\delta}, y_i^{trans}\right) = e^{y_i^{trans}} \cdot \tilde{\delta} \tag{26}$$

transformed value (y_i^{trans}); untransformed value (y_i); median sign ($\tilde{\delta}$)

Experiments without target transformation were performed as well but resulted in poorer model quality.

Model Fit

For feature selection, stepwise regression according to section 1.3.5 was applied.
As regressor, scikit-learns `LinearRegressor` was used. The resulting model equation including all parameters for transformation and scaling was stored in a text file.

Model Assessment

To assess model fit, three different plots were used:

- Predicted versus calculated values
- Residuals versus index
- Histogram of residuals

Moreover, to evaluate model stability, baselines of 1000 randomized conditions within the limits for laser energy and gate delay of the corresponding dataset were plotted and analysed.
Additionally, the modelled baselines were compared to the baselines of regression and the measured median spectra of each condition.

3.4.2.1 Low-alloyed steel (SUS-1R)

To calculate the common polynomial degree of the baseline, baseline regression was applied to all measured spectra of the dataset. An overview of the fitted baselines and the distribution of polynomial degrees for each condition can be seen in Figure 20. As seen visually and by calculation, the median common degree is three. So, the baselines of each replicate constrained to a degree of three were calculated and visualised in Figure 21.
For each condition, at least a few baselines could be calculated.

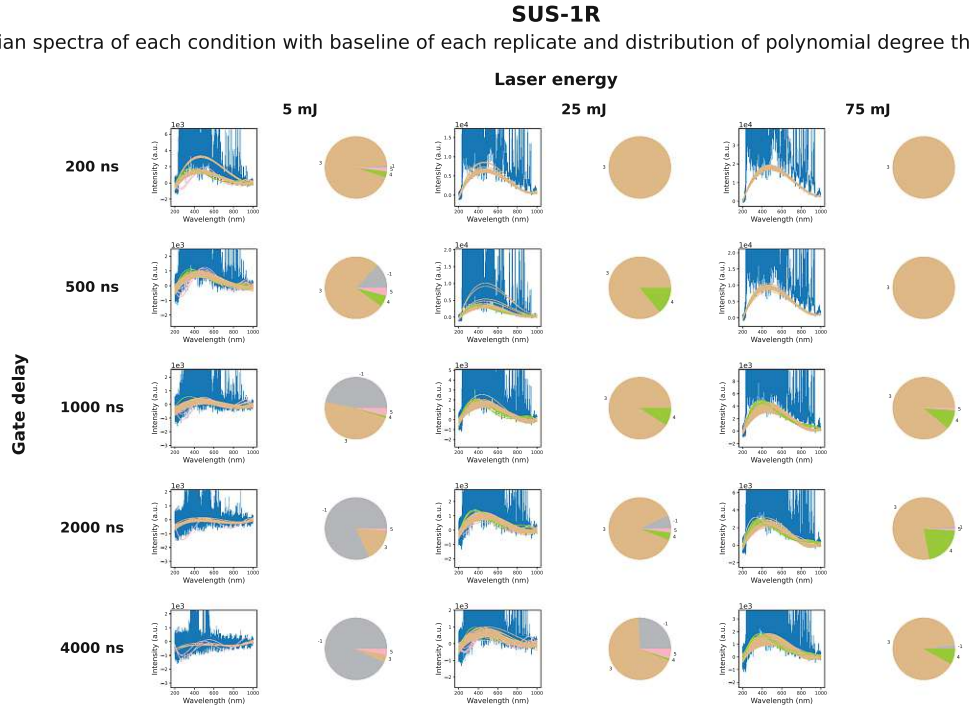


Figure 20: Low-alloyed steel (SUS-1R): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition with baselines (limited by $3 \leq p \leq 5$) of each replicate and pie plots representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 3: beige; 4: green; 5: pink) throughout replicates

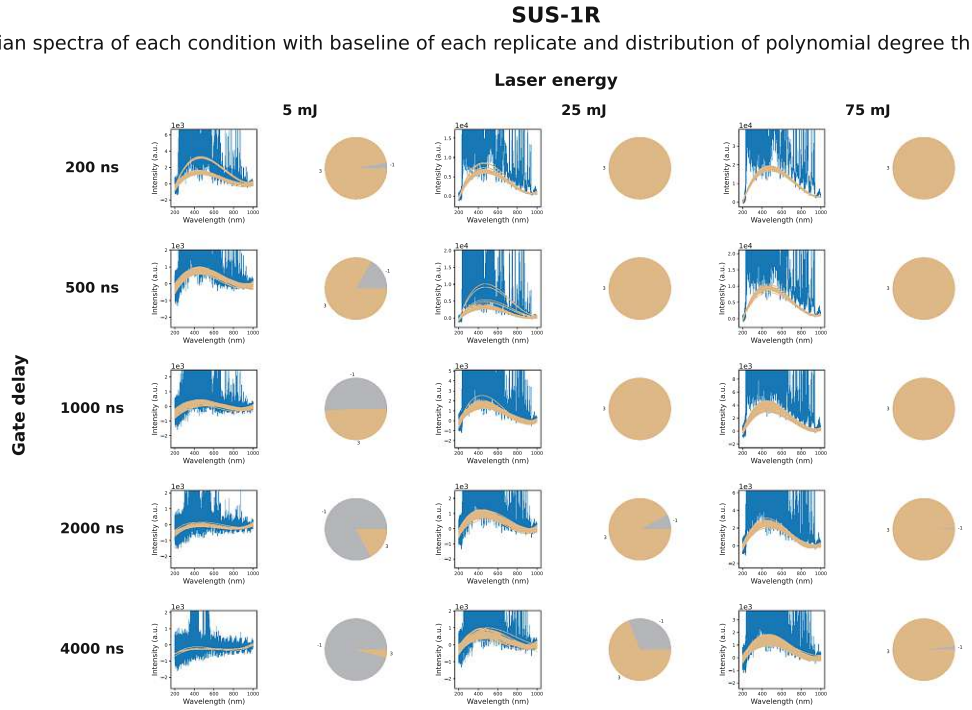


Figure 21: Low-alloyed steel (SUS-1R): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition with baselines (limited by $p = 3$) of each replicate and pie plots representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 3: beige) throughout replicates

Subjecting the data of coefficients in dependency of gate delay and laser energy to the modelling pipeline resulted in the output model equation and F-value for each coefficient (Table 4).

Table 4: Low-alloyed steel (SUS-1R): overview of baseline coefficient model score (F-value) and model equations (numeric values refer to transformed inputs)

Coefficient	F-value	Model equation
a_0	85.44	$a_0(GD, LE) = 1.39 \cdot LE^2 - 0.87 \cdot LE \cdot GD$
a_1	1173.92	$a_1(GD, LE) = -1.36 \cdot GD + 1.03 \cdot LE \cdot GD$
a_2	105.21	$a_2(GD, LE) = -1.33 \cdot GD + 0.98 \cdot LE \cdot GD$
a_3	377.97	$a_3(GD, LE) = 1.44 \cdot LE^2 - 0.99 \cdot LE \cdot GD$

Considering the assessment plots in Figure 22, it can be concluded that the model fits quite well due to normally distributed residuals and well-aligned predicted vs. calculated plots. The skewed residual histogram of coefficient a_2 was accepted since the other assessment plots looked reasonable. Moreover, the model showed stable behaviour in Figure 23 and the final simulated baselines fit in Figure 24 fitted well to the data. For the condition of 200 ns gate delay and 25 mJ laser energy, a constant offset was observed. This was accepted since the offset is constant across the entire wavelength range.

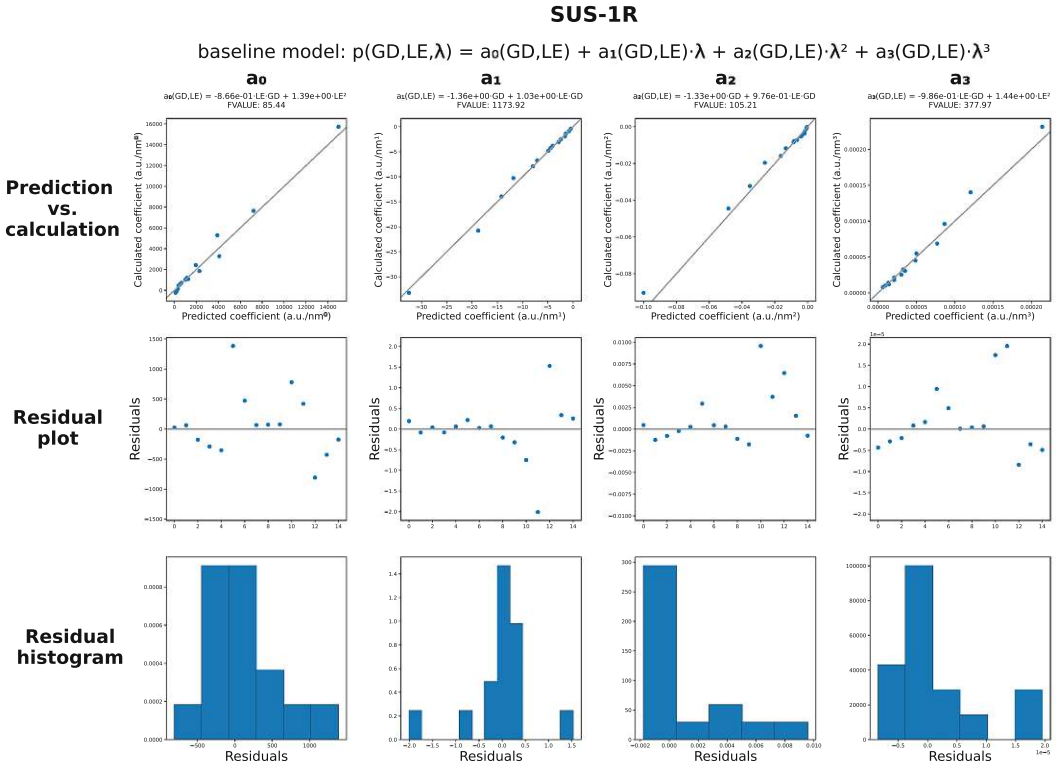


Figure 22: Low-alloyed steel (SUS-1R): assessment plots (prediction vs. calculated, residual plot, histogram of residuals) for each baseline coefficient model

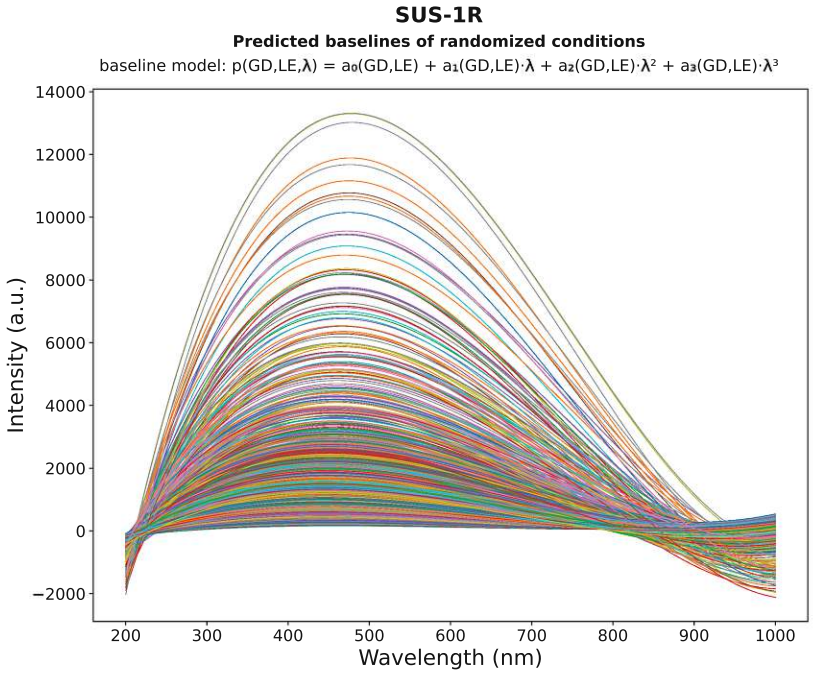


Figure 23: Low-alloyed steel (SUS-1R): simulated baselines of 1000 randomized conditions (different color for each condition) within the ranges of the dataset according to Table 2

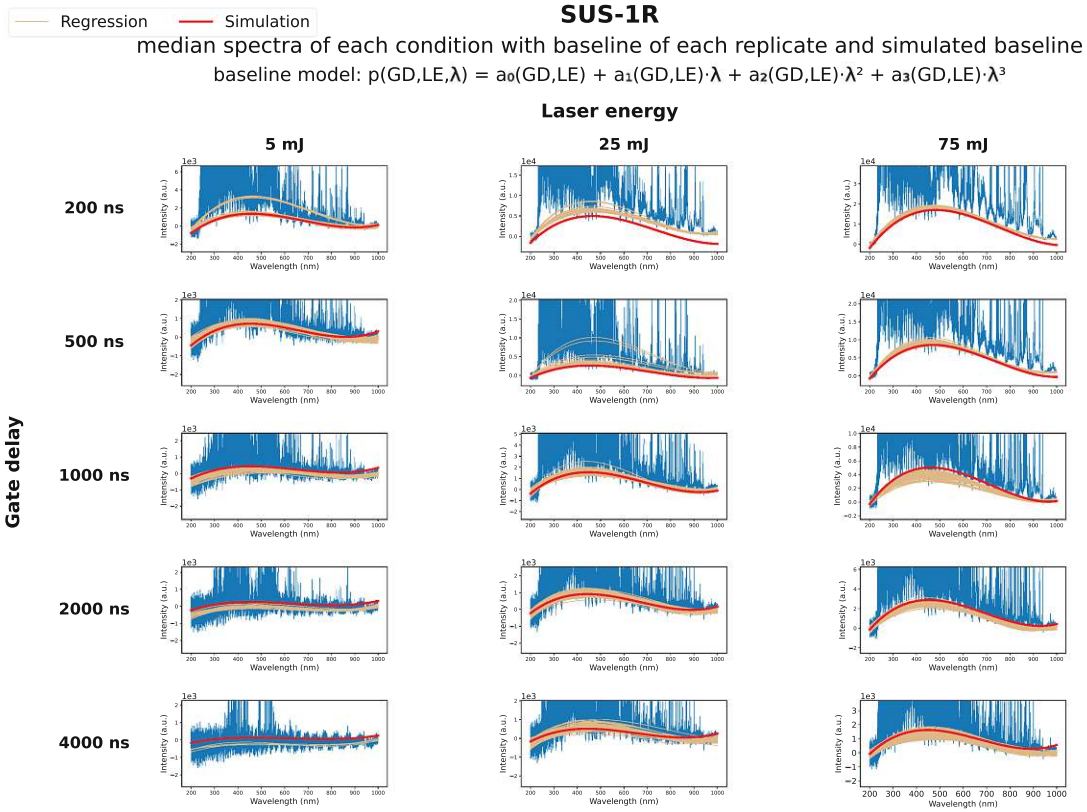


Figure 24: Low-alloyed steel (SUS-1R): median spectra (blue) of each condition with a fitted baseline for each replicate (beige) and simulated baseline (red); Plots: I (a.u.) vs. λ (nm)

3.4.2.2 Aluminium alloy (ERM-EB316)

Limiting the baseline degree between three and five, the baseline fit for all replicates of each condition yielded a common degree of three. However, the baseline fit is oriented on the provided pivot points and not on visual fit quality. For baselines of polynomial degree three, underfitting compared to polynomials of degree four is seen Figure 25. That's why it was opted for a common polynomial degree of four. Figure 26, shows that a baseline fit was possible for each condition.
The calculated model equation of each baseline coefficient and the corresponding F-value are given in Table 5. For two coefficients (a_1, a_3) no model fit with significant coefficients was possible.

Table 5: Aluminium alloy (ERM-EB316): overview of baseline coefficient model score (F-value) and model equations (numeric values refer to transformed inputs)

Coefficient	F-value	Model equation
a_0	165.45	$a_0(GD, LE) = 1.13 \cdot LE^2 - 0.82 \cdot LE \cdot GD$
a_1	-	-
a_2	197.29	$a_2(GD, LE) = -0.97 \cdot GD + 1.77 \cdot LE^2 - 2.08 \cdot LE \cdot GD$
a_3	-	-
a_4	82.91	$a_4(GD, LE) = 1.56 \cdot GD^2 + 2.03 \cdot LE - 2.76 \cdot LE \cdot GD$

The assessment plots of Figure 27 are acceptable for all coefficients. The stability of the model (see Figure 28) seems good as well. Comparing the simulated baselines to measured data and regressed baselines in Figure 29, a slight offset is visible. But, this offset is acceptable due to its consistency across the wavelength range.

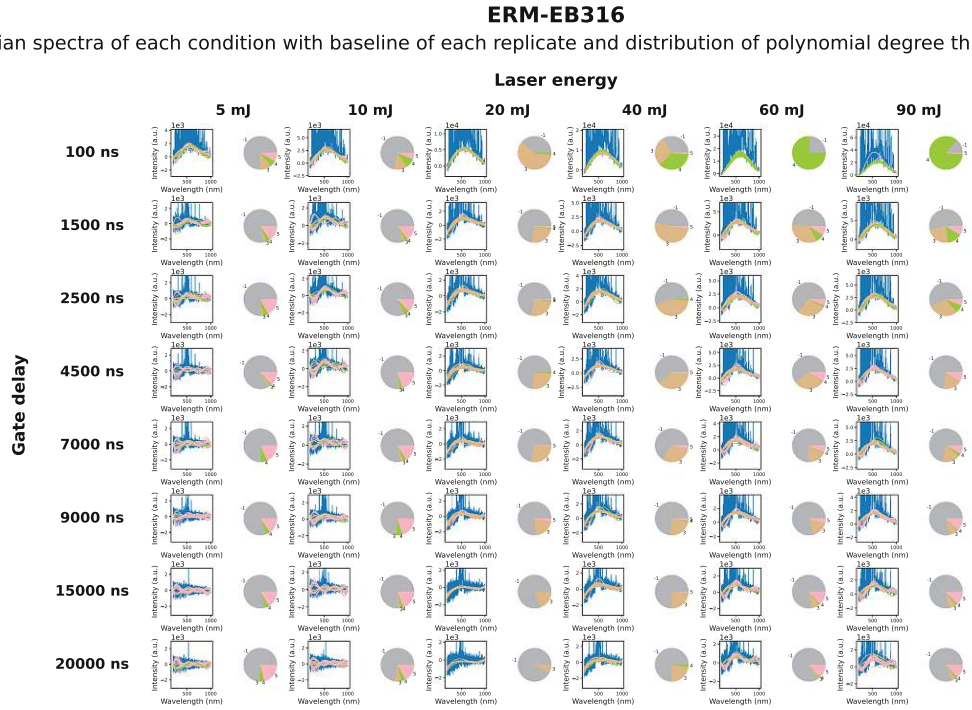


Figure 25: Aluminium alloy (ERM-EB316): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition with baselines (limited by $3 \leq p \leq 5$) of each replicate and pie plots representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 3: beige; 4: green; 5: pink) throughout replicates

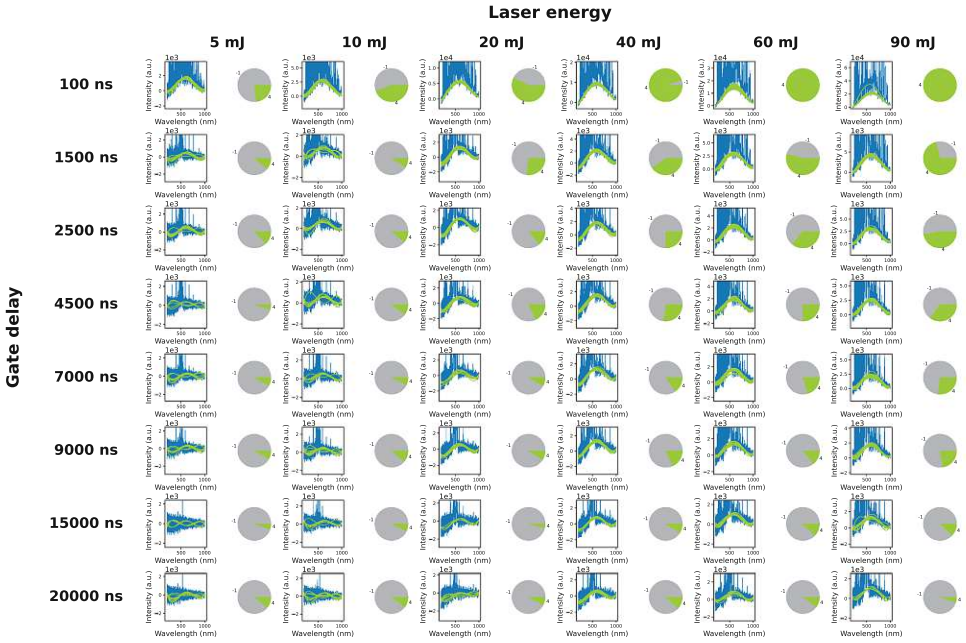


Figure 26: Aluminium alloy (ERM-EB316): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition with baselines (limited by $p = 4$) of each replicate and pie plots representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 4: green) throughout replicates

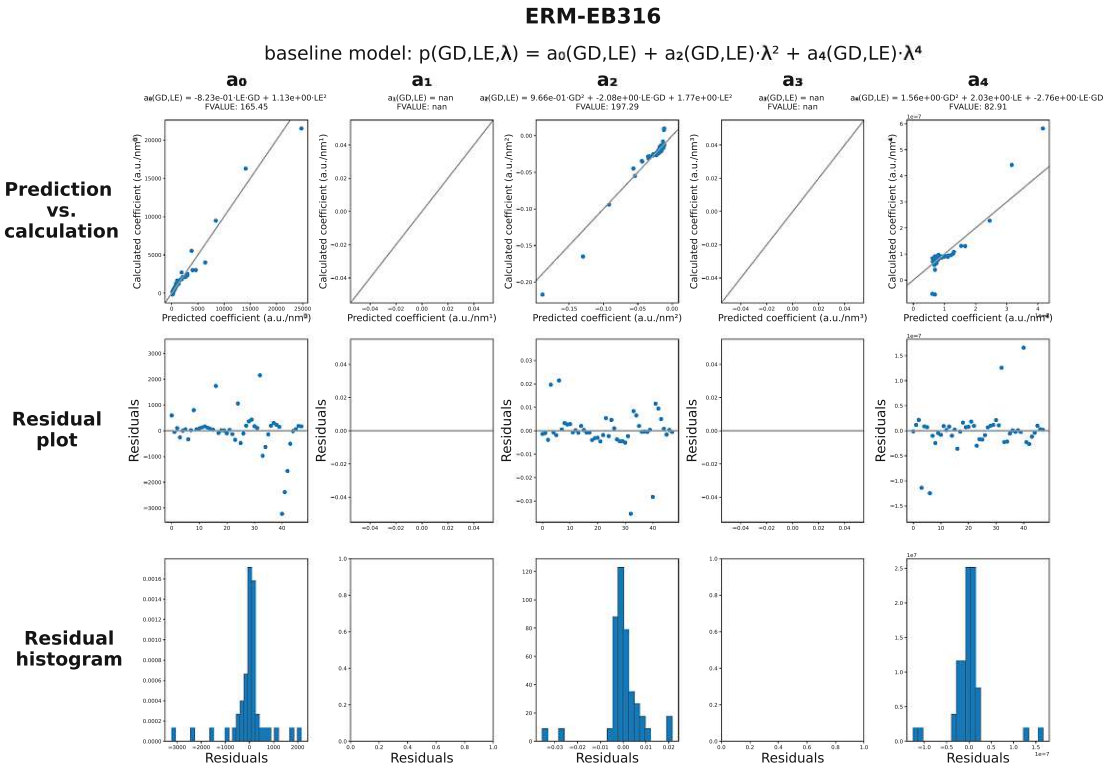


Figure 27: Aluminium alloy (ERM-EB316): assessment plots (prediction vs. calculated, residual plot, histogram of residuals) for each baseline coefficient model, whereas empty plots are displayed if no model could be fitted

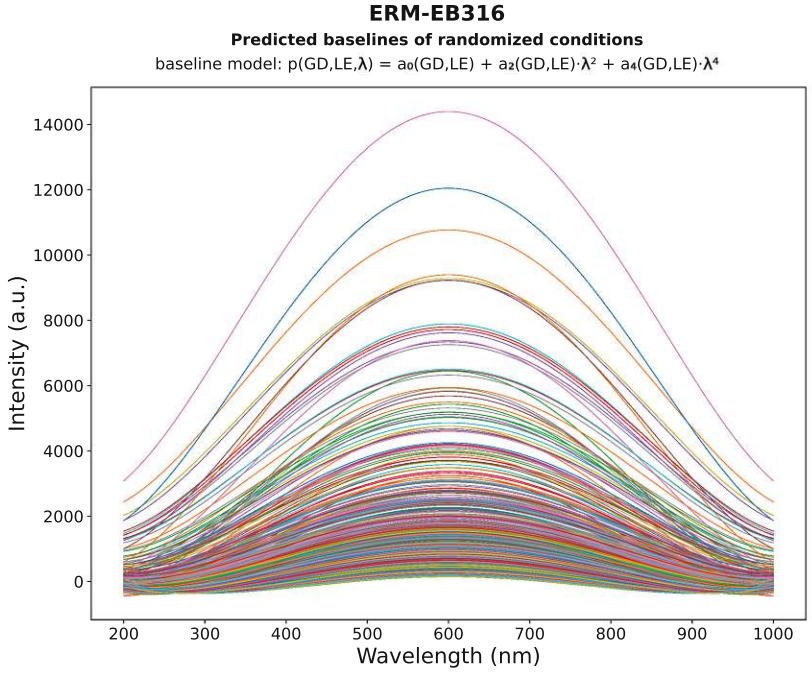


Figure 28: Aluminium alloy (ERM-EB316): simulated baselines of 1000 randomized conditions (different color for each condition) within the ranges of the dataset according to Table 2

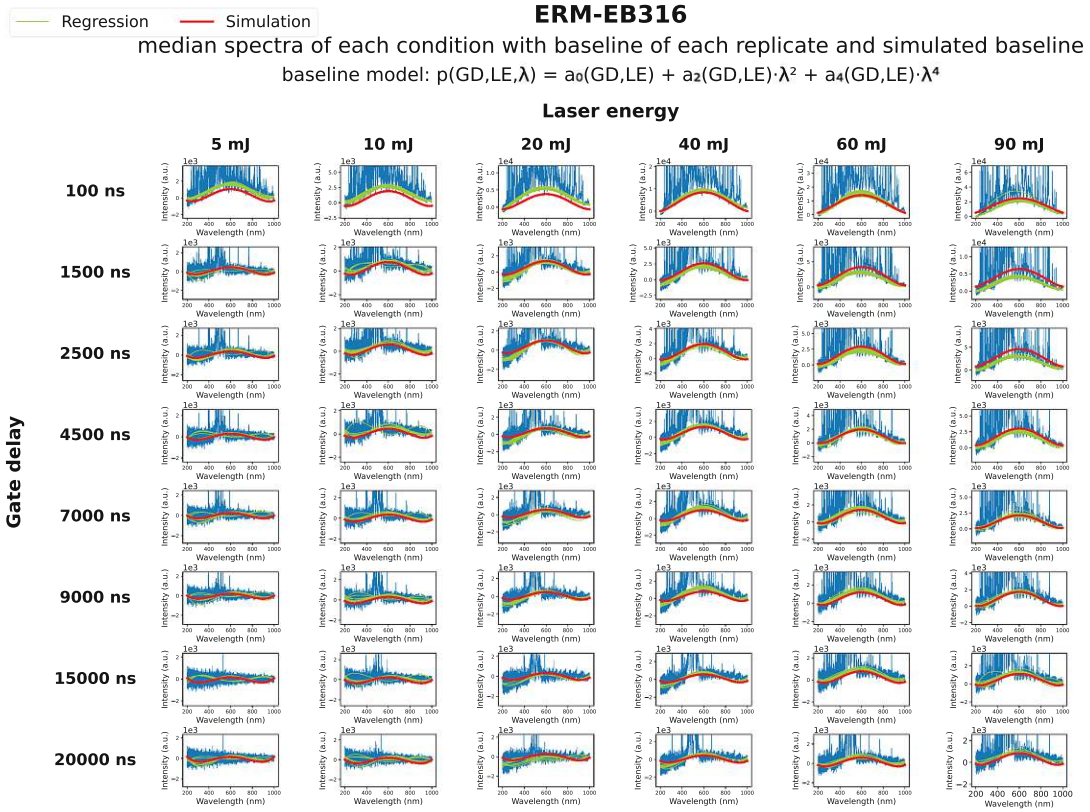


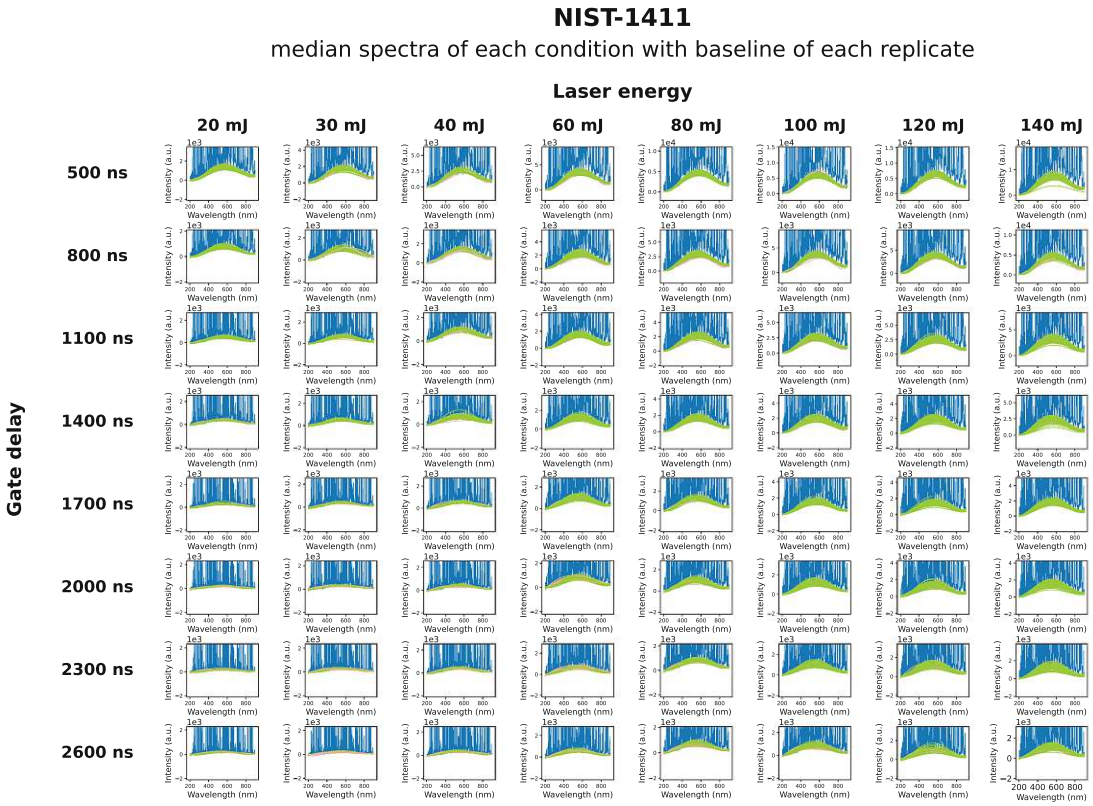
Figure 29: Aluminium alloy (ERM-EB316): median spectra (blue) of each condition with a fitted baseline for each replicate (green) and simulated baseline (red); Plots: I (a.u.) vs. λ (nm)

3.4.2.3 Borosilicate glass (NIST-1411)

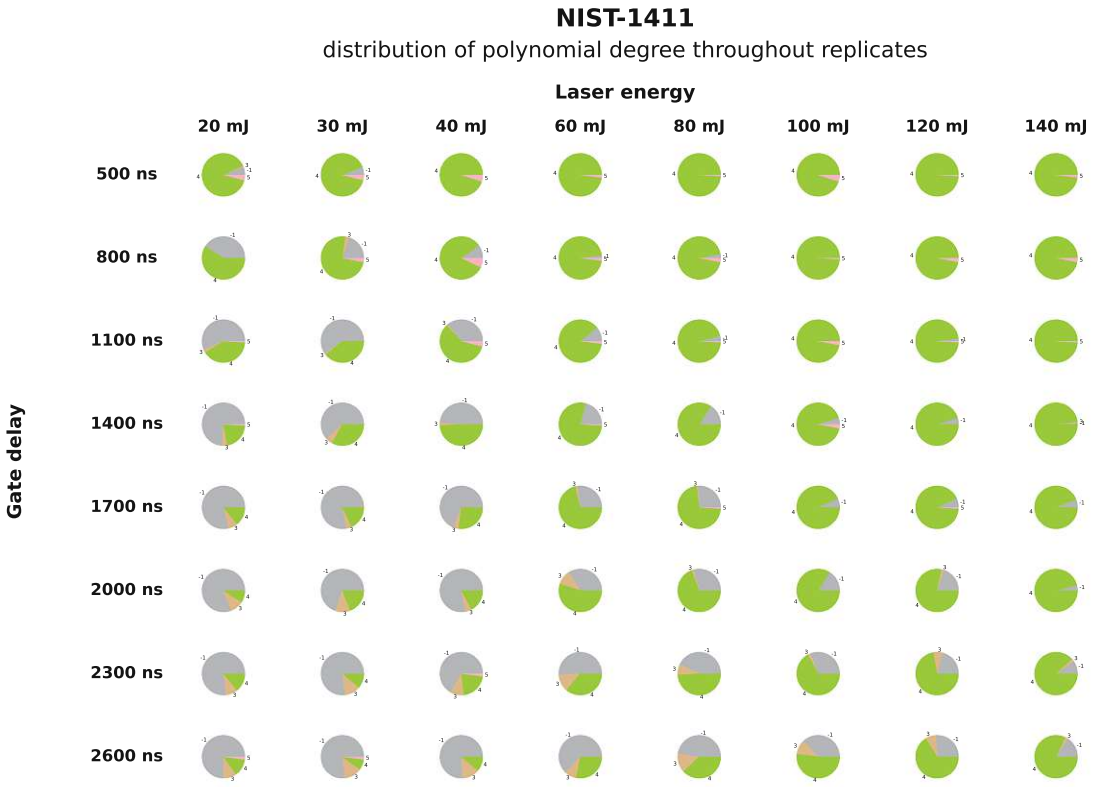
According to Figure 30 and calculation, the common polynomial degree of the baseline was determined as four. Figure 31 also shows that a baseline fit was possible for all conditions.
Table 6 shows the F-value and model equation for each baseline coefficient. For two coefficients (a_1,a_3), no model fit with significant coefficients was possible. The model assessment plots in Figure 32 look good for all coefficients. Moreover, the model appears stable according to Figure 33 and the simulated baselines fit to measured data (see Figure 34).

Table 6: Borosilicate glass (NIST-1411): overview of baseline coefficient model score (F-value) and model equations (numeric values refer to transformed inputs)

Coefficient	F-value	Model equation
a_0	1669.20	$a_0(GD, LE) = 2020 \cdot GD - 2020 \cdot GD^2 - 57.1 \cdot LE + 57.8 \cdot LE^2$
a_1	-	-
a_2	1416.36	$a_2(GD, LE) = 2340 \cdot GD - 2340 \cdot GD^2 + 0.69 LE^2$
a_3	-	-
a_4	1084.89	$a_4(GD, LE) = 2520 \cdot GD - 2520 \cdot GD^2 + 0.67 \cdot LE^2$

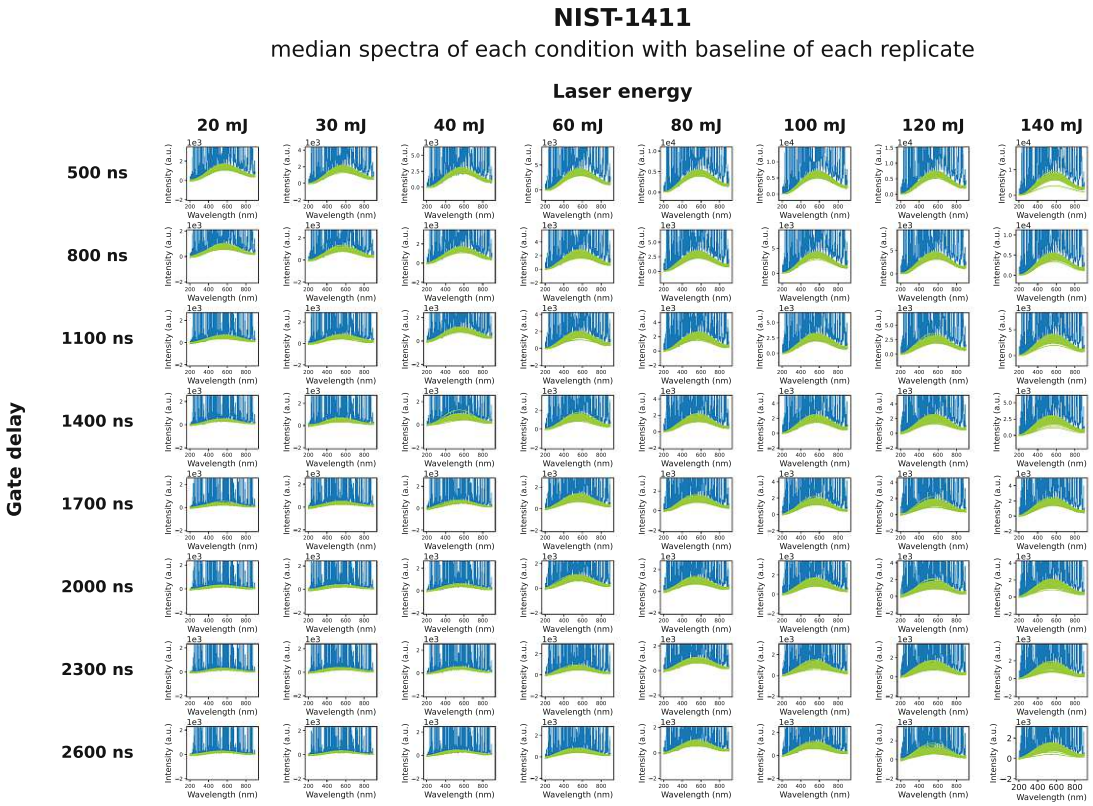


(a) Median spectra of each condition with baselines of each replicate

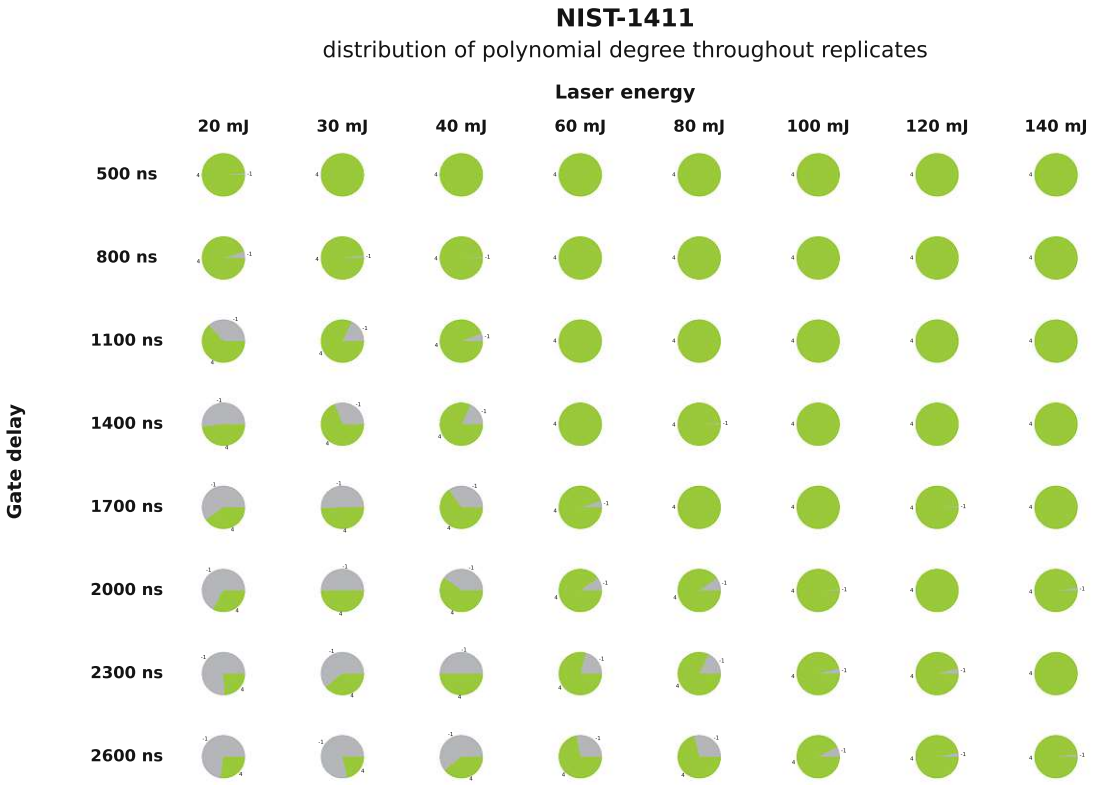


(b) pie plots representing the distribution of polynomial degrees of fitted baselines of (a)

Figure 30: Borosilicate glass (NIST-1411): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition (blue) with baselines (limited by $3 \leq p \leq 5$) of each replicate (a) and pie plots (b) representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 3: beige; 4: green; 5: pink) throughout replicates



(a) Median spectra of each condition with baselines of each replicate



(b) pie plots representing the distribution of polynomial degrees of fitted baselines of (a)

Figure 31: Borosilicate glass (NIST-1411): median spectra (blue; $\tilde{I}_c$ (a.u.) vs. λ (nm)) of each condition with baselines (limited by $p = 4$) of each replicate (a) and pie plots (b) representing the distribution of polynomial degrees (-1(=no fit): grey: no fit; 4: green) throughout replicates

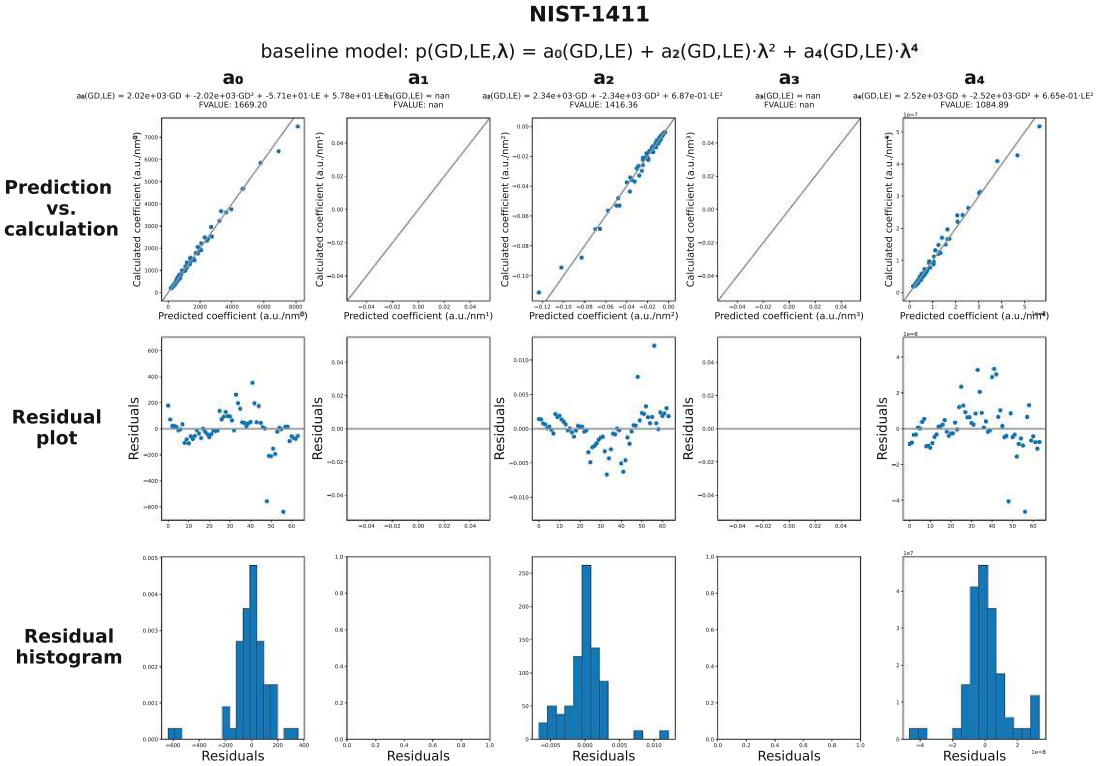


Figure 32: Borosilicate glass (NIST-1411): assessment plots (prediction vs. calculated, residual plot, histogram of residuals) for each baseline coefficient model, whereas empty plots are displayed if no model could be fitted

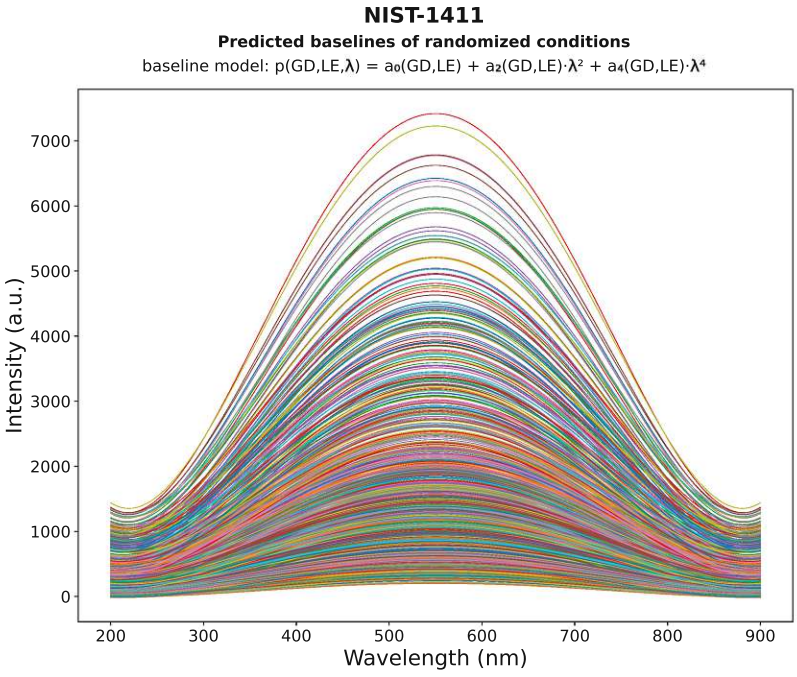


Figure 33: Borosilicate glass (NIST-1411): simulated baselines of 1000 randomized conditions (different color for each condition) within the ranges of the dataset according to Table 2

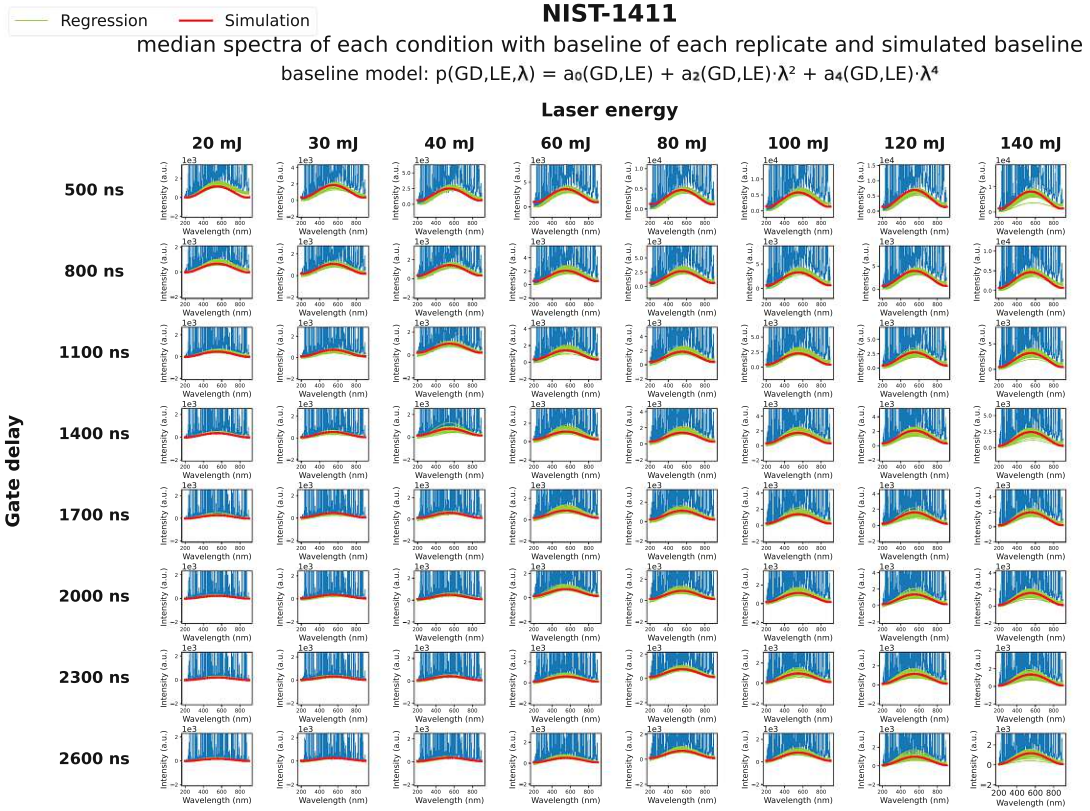


Figure 34: Borosilicate glass (NIST-1411): median spectra (blue) of each condition with a fitted baseline for each replicate (green) and simulated baseline (red); Plots: I (a.u.) vs. λ (nm)

3.4.3 Chapter Summary

Baseline regression was successful for all conditions of each material. However, for some conditions, only a few baselines could be fitted. This resulted from the fact that baselines of the same degree should be fitted for all conditions and for condition rather off from the average, only a few spectra were able to be fitted. Moreover, it was found, that the baselines vary significantly for each material. Regarding baseline simulation, the predicted baselines were satisfying. Still, a constant offset over the wavelength range was visible for some conditions of materials. Since this only affects the intercept of the simulated spectrum, it is acceptable. A summary of the optimised baseline regression settings and the polynomial degree of the baselines of each dataset is given in Table 7.

Table 7: Overview of optimised baseline regression parameters and baseline degree for low-alloyed steel (SUS-1R), aluminium alloy (ERM-EB316) and borosilicate glass (NIST-1411)

Parameter	SUS-1R	ERM-EB316	NIST-1411
window width of the minima filter (nm)	40	40	40
detailed wavelength range (nm)	20	20	20
detail factor	10	10	10
reverse direction	true	false	false
endpoint-weight	10000	10000	1
polynomial degree	3	4	4

3.5 Signal

Prior to line modelling, the present elemental lines had to be identified and the corresponding data points had to be extracted to model the lines. Concerning modelling, each elemental line was modelled separately and experiments were performed to model the line profile or just the line intensity. The models for the lines were based on polynomials just as the baseline models.
Subsequently, the steps of line identification and modelling are discussed in detail. However, the model predictions will not be included in the modelling section and will be discussed as part of the total spectrum simulation.

3.5.1 Line Identification

Database

To get reference values for elemental lines, the AtomTrace LIBS database and the ImageLab LIBS database were used. Both datasets were merged into one removing any duplicates. In total, 2337 lines were included in the database.

Identification

Given the fact that LIBS spectra are made up of multiple lines for each element with varying signal intensity alongside noise, which can overlap with each other as well, line identification results in major challenges when designing an automated algorithm. In this thesis, a comb filter algorithm based on artificial triangle template peaks was used in combination with a threshold signal to distinguish elemental lines from noise. The used algorithm was developed by Zuzana Gajarska and was a prototype at the point of usage in the thesis.
The detailed steps of the algorithm are given below:

1. Application of a median filter to build a baseline
2. Calculation of standard deviation of the baseline based on neighbouring points
3. Calculation of a threshold line by summation of median and standard deviation
4. Identification of lines with the comb algorithm
   (a) Line identification based on database reference values corrected by a given line shift parameter. Only elements present according to Table 3 were identified.
   (b) Fitting a peak template with a defined peak half width (in nm) to identify data points corresponding to the peak
5. Removal of lines below the threshold

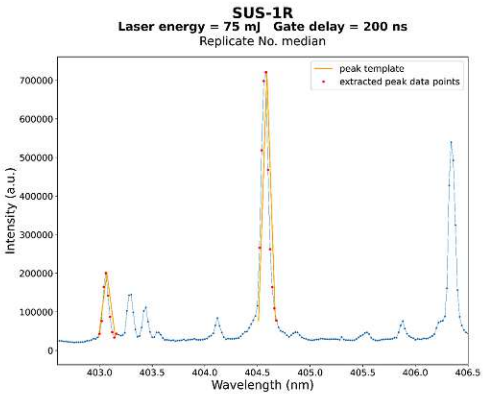
Overall, the following parameters had to be optimised for each dataset:

- window width of the median filter (WW_{med}) in nm
- window width of the standard deviation filter (WW_{std}) in nm
- peak half width (PHW) in nm
- line shift (LS) in nm

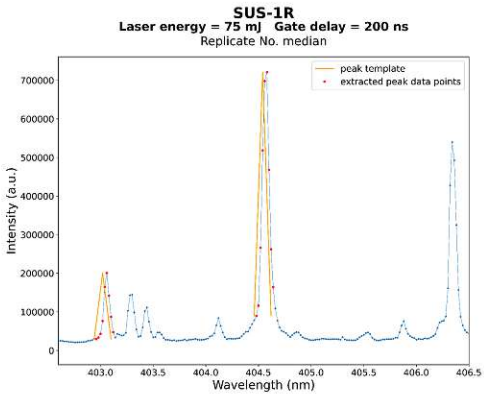
In the following, the optimisation process will be discussed for each dataset.

3.5.1.1 Low-alloyed steel (SUS-1R)

At first, suggested settings by Zuzana Gajarska were used. Therefore, the line with maximum intensity of the majority components (Fe I 404.58 nm) and a smaller line nearby (Mn I 403.06 nm) in the condition with maximum laser energy and minimum gate delay was used to analyse the fit of the peak template (Figure 35a). Here, it can be seen that the peak template is shifted slightly to the right. However, reducing the line shift parameter did not improve the fit (see Figure 35b). Consequently, the default settings were kept as final line identification settings.
From 334 available lines in the database, 189 lines were successfully identified. No lines for sulphur could be identified.



(a) WW_{med}: 12 nm; WW_{std}: 10 nm;
PHW: 0.07 nm; LS: −0.01 nm

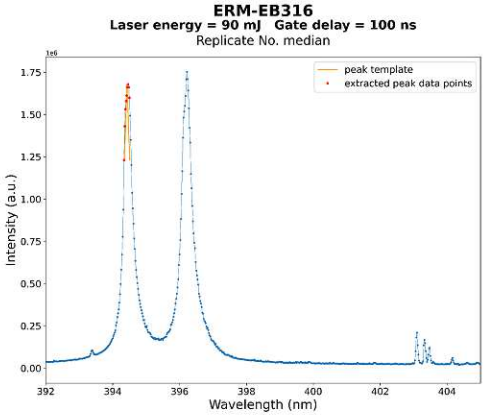


(b) WW_{med}: 12 nm; WW_{std}: 10 nm;
PHW: 0.07 nm; LS: −0.02 nm

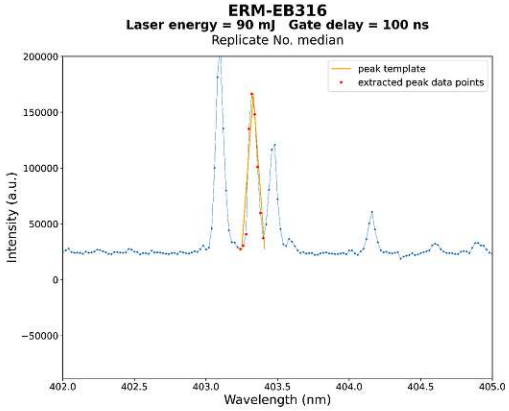
Figure 35: Low-alloyed steel (SUS-1R): median spectra (blue) with fitted peak template (orange) and peak data points (red) for Mn I 403.06 nm and Fe I 404.58 nm - line with default/final (a) and comparative identification settings (b) demonstrated with spectrum measured at 200 ns gate delay and 75 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

3.5.1.2 Aluminium alloy (ERM-EB316)

Just as for low-alloyed steel, the line with the maximum intensity of the main component (Al I 394.41 nm) and a nearby smaller peak (Mn I 403.45 nm) at the condition of minimum gate delay and maximum laser energy were used to analyse the fit of the peak template. With the suggested pre-sets, no major offset is visible but a variation of peak width in dependency of intensity was detected since the peak template with a fixed PHW either fits high-intensity peaks or less strong lines. By the point of writing this thesis, the algorithm was not capable of adjusting to changing peak widths. Consequently, this error was taken into account and it was opted to match the peak width of less intense lines rather than full-scale lines since the majority of lines are of lower intensity. So, the default settings were applied as final line identification settings.
From 237 available lines in the database, 141 lines were found by the algorithm.



(a) line: Al I 394.41 nm
WW_{med}: 15 nm; WW_{std}: 4 nm;
PHW: 0.07 nm; LS: 0.01 nm

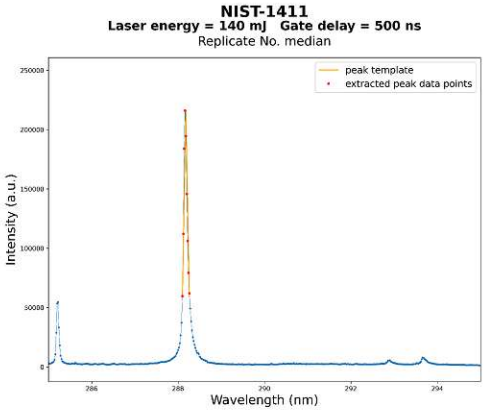


(b) line: Mn I 403.31 nm
WW_{med}: 15 nm; WW_{std}: 4 nm;
PHW: 0.07 nm; LS: 0.01 nm

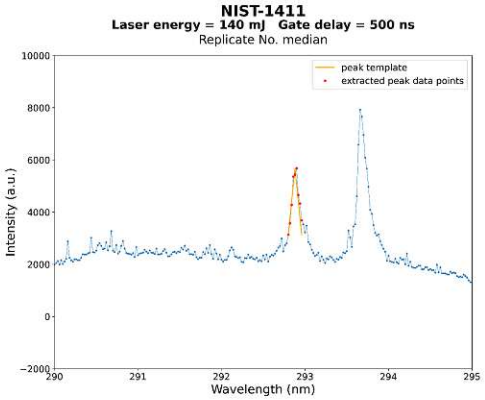
Figure 36: Aluminium alloy (ERM-EB316): median spectra (blue) with fitted peak template (orange) and peak data points (red) for Al I 394.41 nm (a) and Mn I 403.31 nm - line (b) demonstrated with spectrum measured at 100 ns gate delay and 90 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

3.5.1.3 Borosilicate glass (NIST-1411)

For this material, the most intense line of the main component (Si I 228.17 nm) and a smaller line (Mg I 292.88 nm) was used to analyse the fit of the peak template for the condition with minimum gate delay and maximum laser energy. Here, a confident template fit with the suggested settings by Zuzana Gajarska was achieved (Figure 37). Yet, the problem of non-consistent peak width occurred again and it was also decided to keep the settings in favour of smaller lines. From 315 database lines, 197 could be identified.



(a) line: Si I 228.17 nm
WW_{med}: 15 nm; WW_{std}: 4 nm;
PHW: 0.07 nm; LS: 0.01 nm



(b) line: Mg I 292.88 nm
WW_{med}: 15 nm; WW_{std}: 4 nm;
PHW: 0.07 nm; LS: 0.01 nm

Figure 37: Borosilicate glass (NIST-1411): median spectra (blue) with fitted peak template (orange) and peak data points (red) for Si I 228.17 nm (a) and Mg I 292.88 nm - line (b) demonstrated with spectrum measured at 500 ns gate delay and 140 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

3.5.2 Modelling

Workflow

To build a model for each line in dependency of gate delay and laser energy, the same line had to be identified in all conditions. Therefore, line identification was performed for all replicates at the condition of lowest gate delay and highest laser energy. Lines present in more than 50 % of the replicates were then preselected. These lines were then identified in the median spectra of all conditions independently of their intensity concerning the given threshold. So, the data points of all identified lines in dependency on gate delay and laser energy were received and subjected to build the models.

Profile Model

For each line, a Pseudo-Voigt profile (Equation 6) was fitted with `PseudoVoigtModel(...)` of the `lmfit` library. This was included in the method `Spectrum.Lines.identify_spectra_lines(...)`. Based on the median spectrum with minimum gate delay and maximum laser energy, lines with no proper fit were excluded. The criteria for a successful fit were a Spearman's rank coefficient > 0.7 of the extracted data points and the corresponding points of the fitted profile and a difference of the centre of the fitted profile $\lambda_{0\,fit}$ and the centre of the extracted line data points λ_{0data} (point with maximum intensity) < 0.05 nm. For other conditions, no lines were discarded to guarantee the values for σ, A and θ are obtained for each line and all conditions.
Next, separate models for σ, A and θ in dependency on gate delay and laser energy were fitted. For the polynomials, the maximum degree was limited to two including cross terms. Throughout preprocessing, the feature matrix except for the intercept column was transformed according to Equation 9 with subsequent scaling (Equation 11). The targets were only transformed by a custom transformer as described in section 3.4.2 and Equation 24-26 if it increased the model quality. For feature selection, a stepwise algorithm with F-value as scoring criteria was applied (section 1.3.5).

After model fit, models with an F-value < 30 for A or an F-value < 10 for σ were discarded. No exclusion criterium for θ was added since the number of models was already significantly reduced by the criteria for A and σ. Optionally, models only made up of terms of gate delay or laser energy could be dropped as well. In general, models dependent on both variables should be favoured since the final simulation should be based on both variables as well. Moreover, just as for baseline modelling, plots (prediction vs. calculation; residuals vs. index, histogram of residuals) were used for visual model assessment.
So, three models according to Equation 27-29 were achieved for each remaining line.

$$\sigma_{Element\ Charge\ Wavelength}(GD, LE) = a_0 + a_1 \cdot GD + a_2 \cdot GD^2 + a_3 \cdot GD \cdot LE + a_4 + a_5 \cdot LE + \cdot LE^2 \quad (27)$$

standard deviation (σ); gate delay (GD); laser energy (LE); coefficient (a)

$$A_{Element\ Charge\ Wavelength}(GD, LE) = a_0 + a_1 \cdot GD + a_2 \cdot GD^2 + a_3 \cdot GD \cdot LE + a_4 + a_5 \cdot LE + \cdot LE^2 \quad (28)$$

peak height (A); gate delay (GD); laser energy (LE); coefficient (a)

$$\theta_{Element\ Charge\ Wavelength}(GD, LE) = a_0 + a_1 \cdot GD + a_2 \cdot GD^2 + a_3 \cdot GD \cdot LE + a_4 + a_5 \cdot LE + \cdot LE^2 \quad (29)$$

shape coefficient (θ); gate delay (GD); laser energy (LE); coefficient (a)

Intensity Model

As a simplification, models for just the peak intensity were generated. In detail, the peak intensity was calculated as the difference between the minimum and maximum intensity values of the extracted data points of an elemental line. Just as for the profile models, the maximum polynomial degree was limited to two and the feature matrix except the intercept column was transformed and scaled. Also, the intensity, as the target variable, was only transformed and scaled if it improved the model. Again, stepwise regression with F-value as scoring parameter was applied (see section 1.3.5).
Moreover, models with an F-value < 30 were discarded and models only depending on laser energy or gate delay could be removed optionally. The same plots as for the profile models were used for visual assessment.
Finally, an intensity model (Equation 30) was returned for each remaining line.

$$I_{Signal;Element\ Charge\ Wavelength}(GD, LE) = a_0 + a_1 \cdot GD + a_2 \cdot GD^2 + a_3 \cdot GD \cdot LE + a_4 + a_5 \cdot LE + \cdot LE^2 \quad (30)$$

intensity (I); gate delay (GD); laser energy (LE); coefficient (a)

Subsequently, the modelling process of each material will be discussed. Supplementary data to the following discussion is provided at `https://zenodo.org/records/10557230`.

3.5.2.1 Low-alloyed steel (SUS-1R)

Profile Model

Figure 38 shows the identified lines with the fitted Pseudo-Voigt profile after the exclusion of misfitted lines (ρ_S<0.7 and $|\lambda_{0fit} - \lambda_{0data}|$ <0.05 nm) for the condition of minimum gate delay and maximum laser energy. Here, it is seen that the fit works quite well for the majority of lines. For overlapping lines, such as Cr I 276.26 nm, the fit only covered the general line structure as a consequence of not sufficiently resolved peak data point identification. From 189 identified lines, 171 were successfully fitted.
Subjecting the fitted profiles to modelling, model assessment plots for each parameter of the Pseudo-Voigt profile (A, σ, θ) for each elemental line were obtained. Figure 40 shows these plots exemplary for Manganese. Here, it can be seen that a successful fit was not possible for all lines. So, the exclusion criteria for A and σ models were applied subsequently. Consequently, lines like Cr II 284.32 nm are omitted. With these criteria, the total number of lines was reduced to 84 lines. Moreover, excluding models only dependent on gate delay or laser energy reduced the available lines to 14. As a result, this measure was not applied to the final simulation.

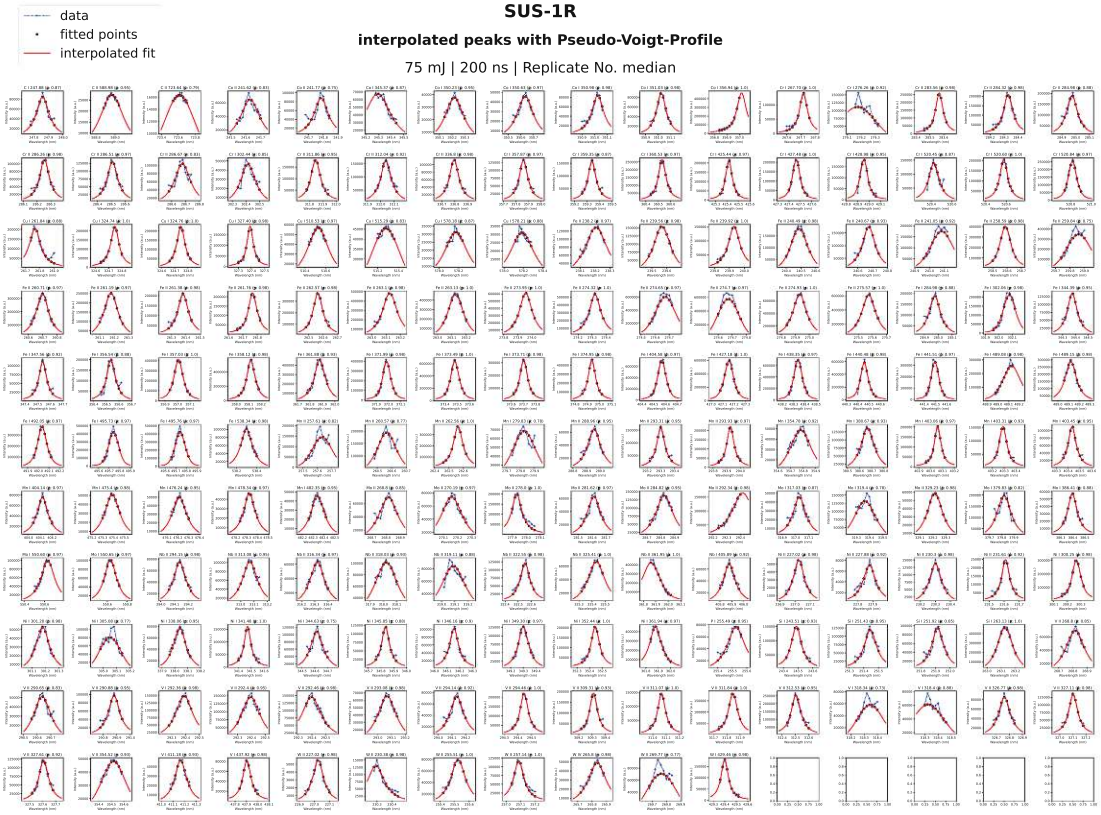


Figure 38: Low-alloyed steel (SUS-1R): identified peak data points (blue) with fitted Pseudo-Voigt profile (red) and fitted points (black) after removal of misfitted lines (ρ_S<0.7 and $|\lambda_{0fit} - \lambda_{0data}|$ <0.05 nm) median spectra at 200 ns gate delay and 75 mJ laser energy; Plots: I (a.u.) vs. λ (nm)

Intensity Model

For intensity models, model assessment plots were obtained as well. Again, Figure 39 shows a model assessment plot exemplary for chromium and the exclusion criterium of F-value < 30 was applied to drop lines like Cr II 540.98 nm. So, from 189 identified lines, 185 were successfully modelled if the models were restricted to depend on both independent variables (GD and LE). One additional model was obtained without this restriction. For the final simulation, only models depending on gate delay and laser energy were selected.

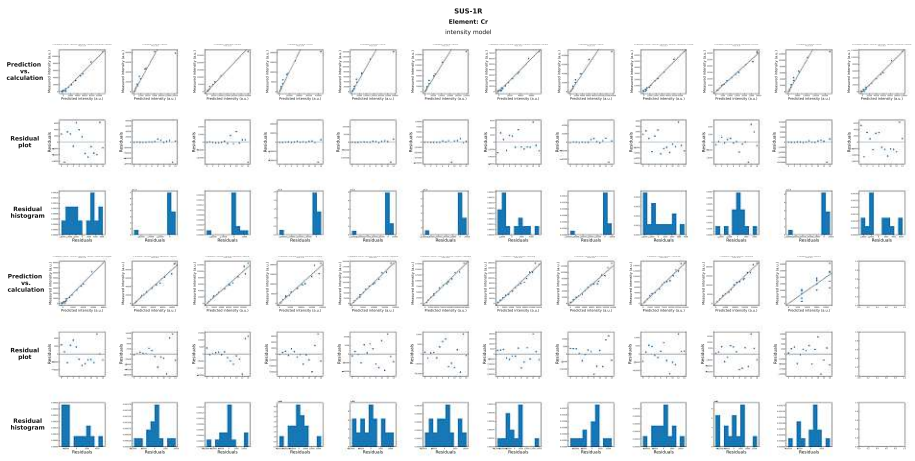
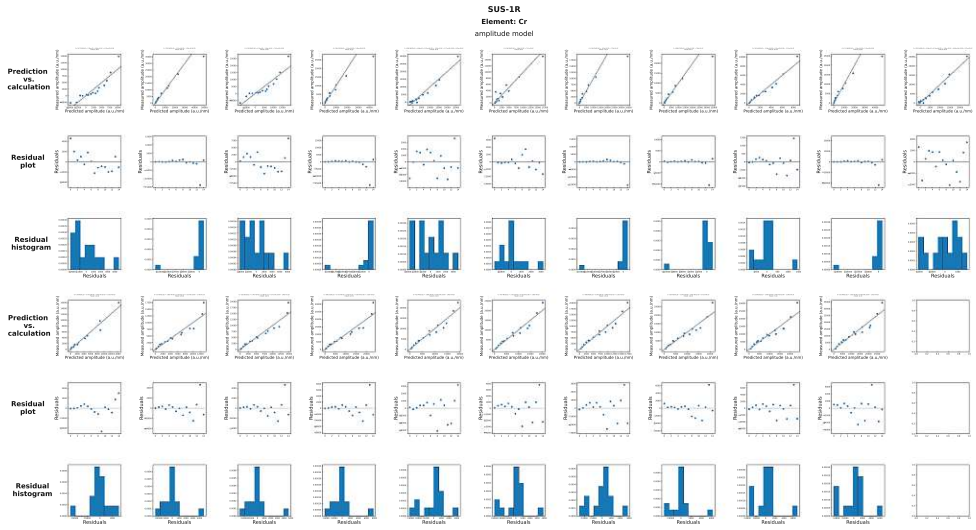
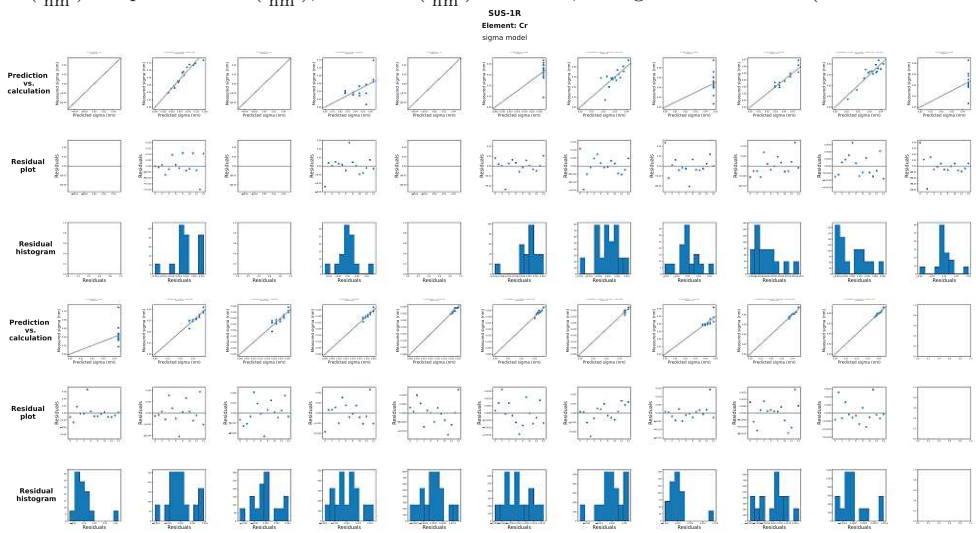


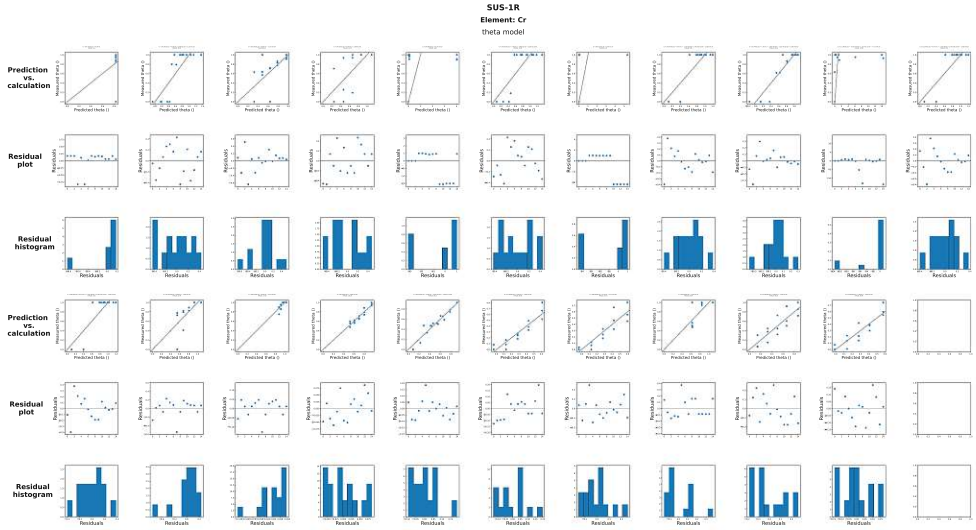
Figure 39: Low-alloyed steel (SUS-1R): assessment plots (measured I (a.u.) vs. predicted I (a.u.), residuals (a.u.) vs. index, histogram of residuals (count vs. residuals (a.u.))) for intensity models of each chromium line prior exclusion of non-sufficient models



(a) A models:
measured A ($\frac{\text{a.u.}}{\text{nm}}$) vs. predicted A ($\frac{\text{a.u.}}{\text{nm}}$); residuals ($\frac{\text{a.u.}}{\text{nm}}$) vs. index; histogram of residuals (count vs. residuals ($\frac{\text{a.u.}}{\text{nm}}$))



(b) σ models:
measured σ (a.u.) vs. predicted σ (a.u.); residuals (a.u.) vs. index; histogram of residuals (count vs. residuals (a.u.))



(c) θ models:
measured θ () vs. predicted θ (); residuals () vs. index; histogram of residuals (count vs. residuals ())

Figure 40: Low-alloyed steel (SUS-1R): assessment plots (prediction vs. calculation, residual plot, histogram of residuals) for A (a), σ (b) and θ models (c) of each chromium line prior exclusion of non-sufficient models, whereas empty plots are displayed if no model could be fitted

3.5.2.2 Aluminium alloy (ERM-EB316)

Profile model

The identified lines with fitted Pseudo-Voigt profiles after the exclusion of misfitted lines for the condition of minimum gate delay and maximum laser energy are visible in Figure 41. From 141 identified lines, 120 were fitted successfully.
As an example, the model assessment plots for all aluminium lines before the exclusion of poor models are given (Figure 42). Given the two exclusion criteria as discussed in section 3.5.2, it can be seen that these criteria do not cover all poor models such as the *A* model for Al I 309.28 nm. Yet, considering the number of models, eliminating models manually is not an option. So, tuning the selection criteria should be improved in the future. Finally, merely 26 lines could be modelled and only 2 lines include laser energy as well as gate delay as independent variables.

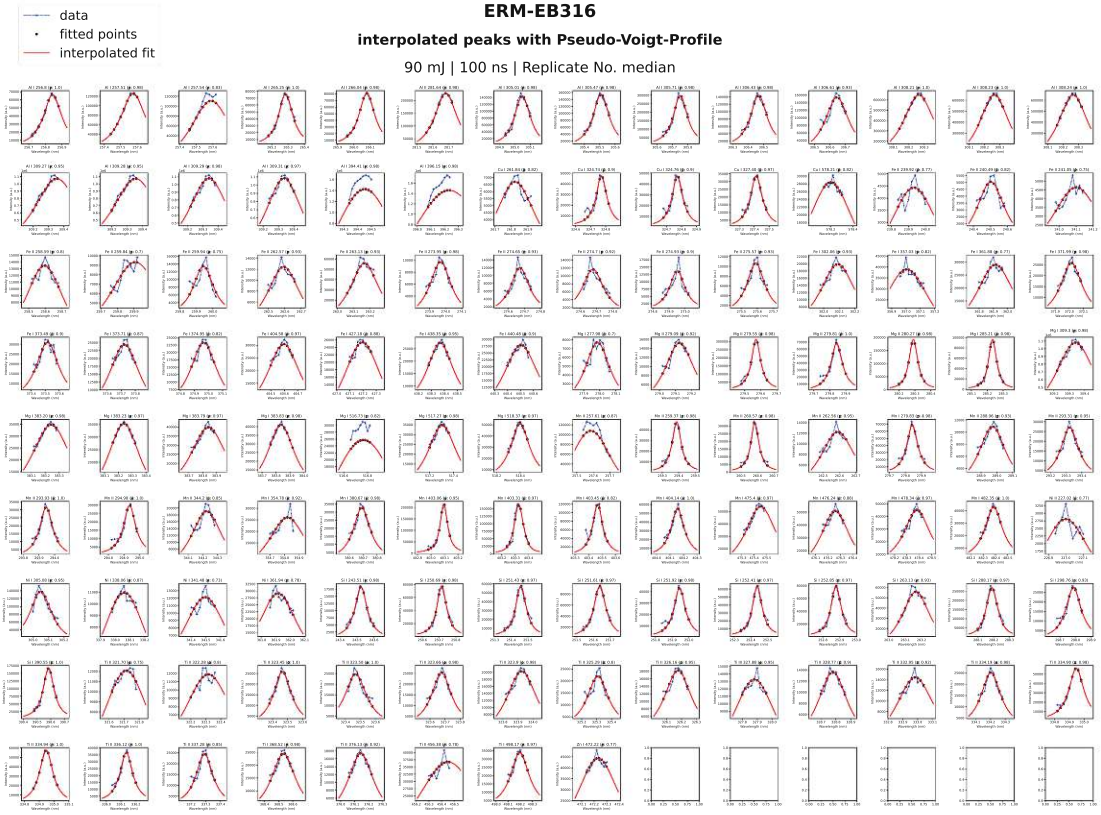
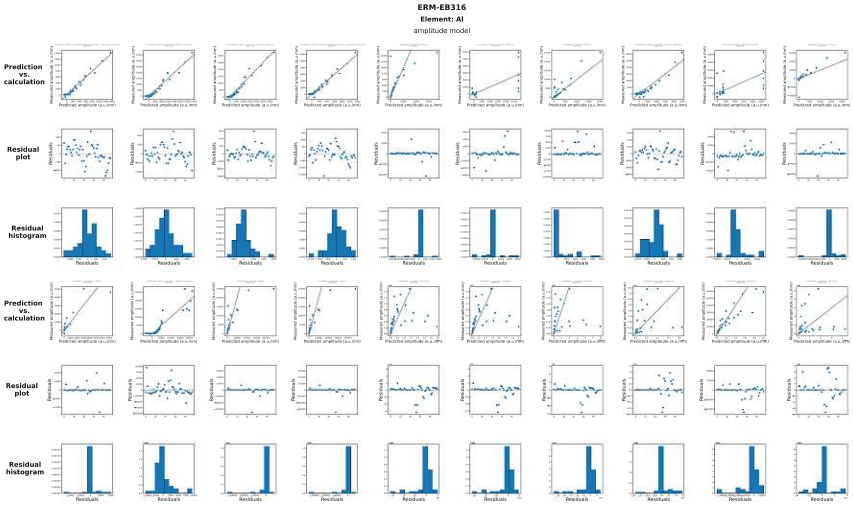
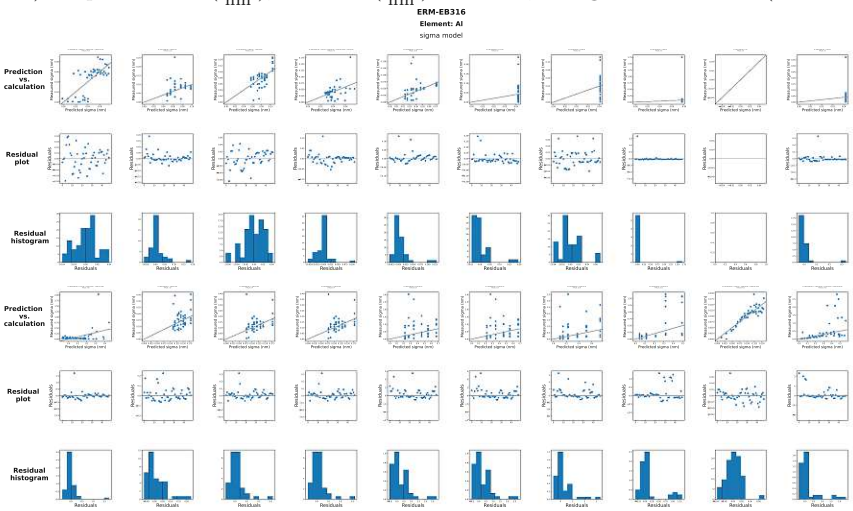


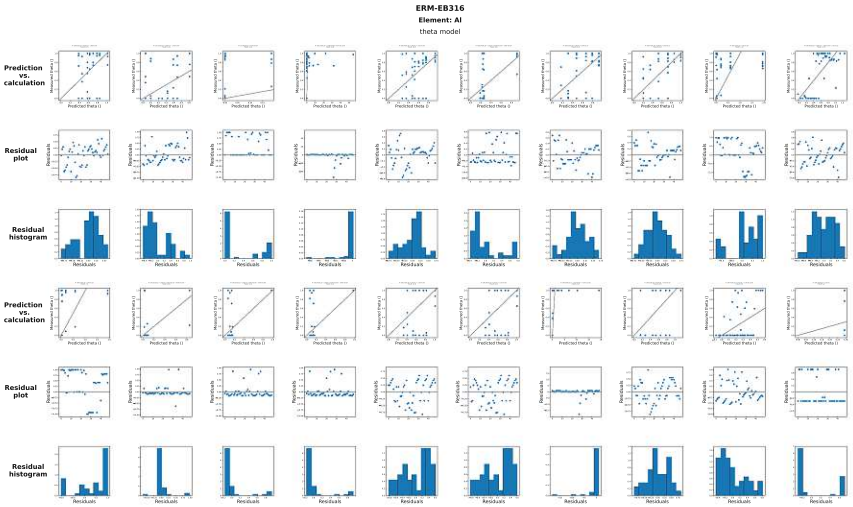
Figure 41: Aluminium alloy (ERM-EB316): identified peak data points (blue) with fitted Pseudo-Voigt profile (red) and fitted points (black) after removal of misfitted lines (ρ_S<0.7 and $|\lambda_{0fit} - \lambda_{0data}|$ <0.05 nm) median spectra at 100 ns gate delay and 90 mJ laser energy, whereas empty plots are displayed if no model could be fitted; Plots: I (a.u.) vs. λ (nm)



(a) A models:
measured A (a.u./nm) vs. predicted A ($\frac{\text{a.u.}}{\text{nm}}$); residuals ($\frac{\text{a.u.}}{\text{nm}}$) vs. index; histogram of residuals (count vs. residuals ($\frac{\text{a.u.}}{\text{nm}}$))



(b) σ models:
measured σ (a.u.) vs. predicted σ (a.u.); residuals (a.u.) vs. index; histogram of residuals (count vs. residuals (a.u.))



(c) θ models:
measured θ () vs. predicted θ (); residuals () vs. index; histogram of residuals (count vs. residuals ())

Figure 42: Aluminium alloy (ERM-EB316): assessment plots (prediction vs. calculation, residual plot, histogram of residuals) for A (a), σ (b) and θ models (c) of each aluminium line prior exclusion of non-sufficient models, whereas empty plots are displayed if no model could be fitted

Intensity model

Figure 43 shows the intensity model assessment plots for all aluminium lines exemplary prior to the exclusion of poor models as discussed in section 3.5.2. From 141 lines, 110 could be modelled successfully. For this dataset, the number of good models is independent from the inclusion or exclusion of models depending on only one independent variable (LE or GD).

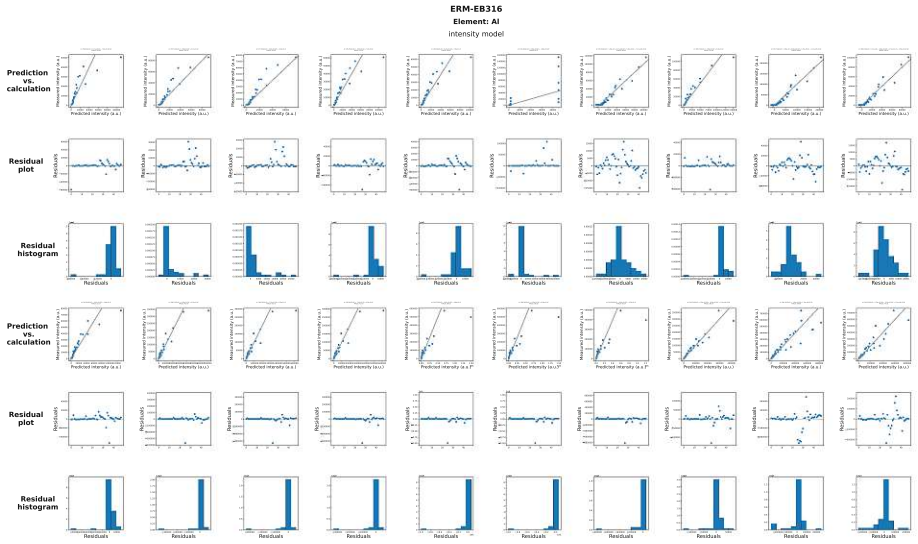


Figure 43: Aluminium alloy (ERM-EB316): assessment plots (measured I (a.u.) vs. predicted I (a.u.), residuals (a.u.) vs. index, histogram of residuals (count vs. residuals (a.u.))) for intensity models of each aluminium line prior exclusion of non-sufficient models

3.5.2.3 Borosilicate glass (NIST-1411)

Profile Model

Figure 44 shows the fitted Pseudo-Voigt profile to the identified lines after the exclusion of weak fits for the condition of minimum gate delay and maximum laser energy. From 197 identified lines, 187 could be fitted successfully. Exemplary, Figure 45a shows the model assessment plots of all silicon lines. For silicon, no bad lines could be identified. Overall, the modelling process worked very well for this dataset and resulted in 169 line profile models and 63 line profile models considering only models including terms of both independent variables(LE and GD). For the final simulation, it was opted to also include the models with single dependency as well.

Intensity Model

Figure 46 shows the intensity model assessment plots of all silicon lines prior to the exclusion of poor models. Overall, 194 lines out of 197 identified lines could be modelled successfully. If only models dependent on laser energy and gate delay should be considered, the number was reduced to 114. Compared to the other two datasets, for the final simulation, models with dependency on only one variable were used due to the huge difference in available lines.

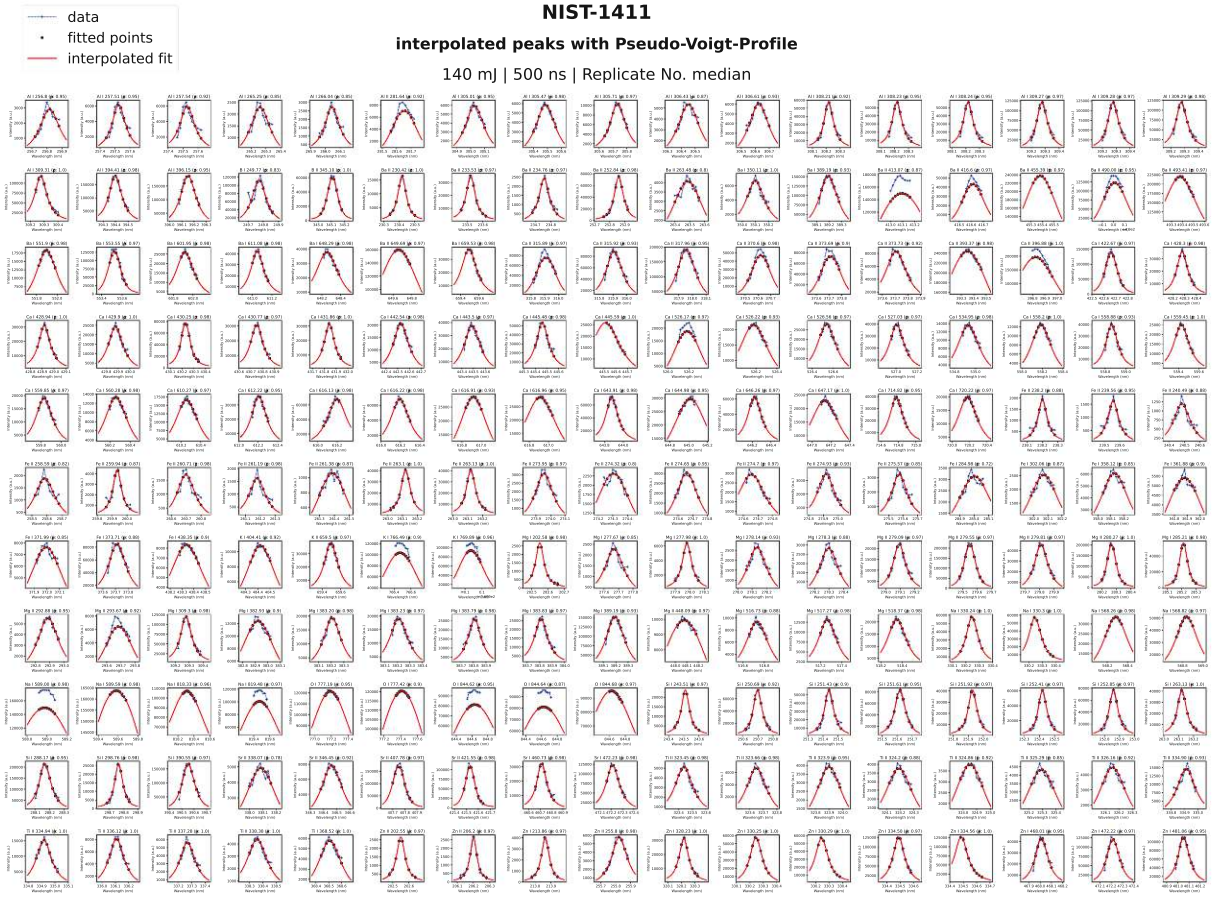
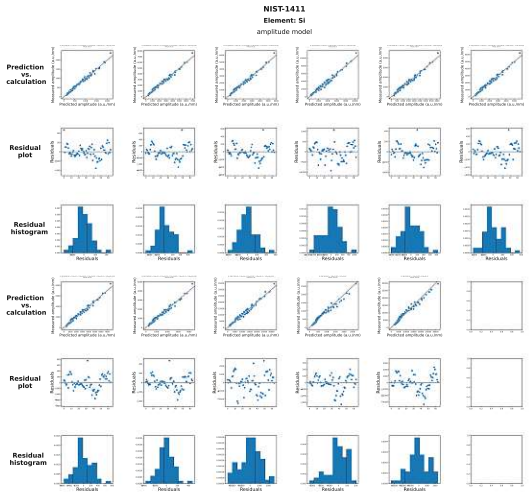
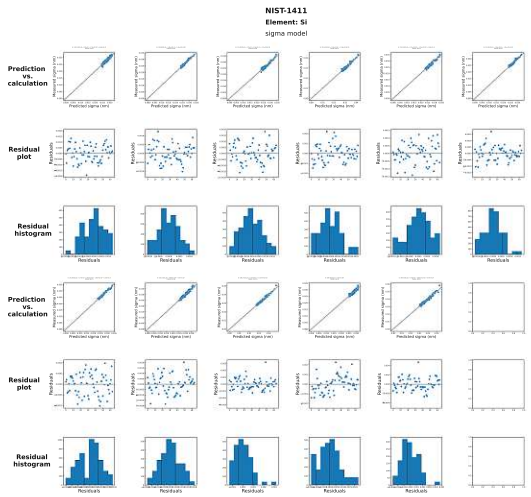


Figure 44: Borosilicate glass (NIST-1411): identified peak data points (blue) with fitted Pseudo-Voigt profile (red) and fitted points (black) after removal of misfitted lines (ρ_S<0.7 and $|\lambda_{0fit} - \lambda_{0data}|$ <0.05 nm) median spectra at 500 ns gate delay and 140 mJ laser energy, whereas empty plots are displayed if no model could be fitted

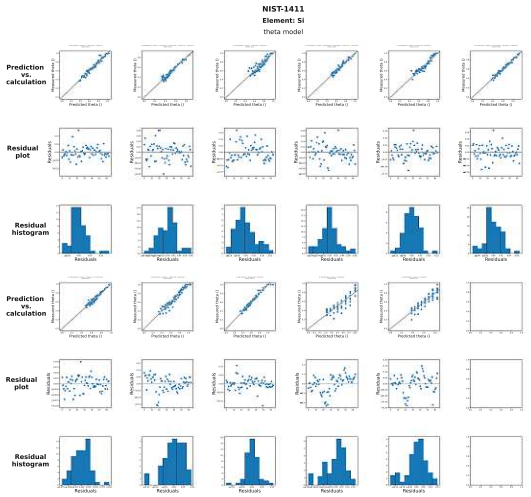


(a) A models:
measured A ($\frac{\text{a.u.}}{\text{nm}}$) vs. predicted A ($\frac{\text{a.u.}}{\text{nm}}$); residuals (a.u./nm) vs. index; histogram of residuals (count vs. residuals ($\frac{\text{a.u.}}{\text{nm}}$))

Figure 45a: Borosilicate glass (NIST-1411): assessment plots (prediction vs. calculation, residual plot, histogram of residuals) for A (a), σ (b) and θ models (c) of each silicon line prior exclusion of non-sufficient models, whereas empty plots are displayed if no model could be fitted



(b) σ models:
measured σ (a.u.) vs. predicted σ (a.u.); residuals (a.u.) vs. index; histogram of residuals (count vs. residuals (a.u.))



(c) θ models:
measured θ () vs. predicted θ (); residuals () vs. index; histogram of residuals (count vs. residuals ())

Figure 45b,c: Borosilicate glass (NIST-1411): assessment plots (prediction vs. calculation, residual plot, histogram of residuals) for A (a), σ (b) and θ models (c) of each silicon line prior exclusion of non-sufficient models, whereas empty plots are displayed if no model could be fitted

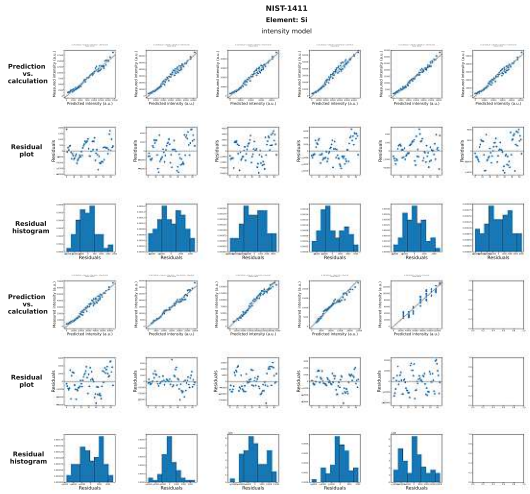


Figure 46: Borosilicate glass (NIST-1411): assessment plots (measured I (a.u.) vs. predicted I (a.u.), residuals (a.u.) vs. index, histogram of residuals (count vs. residuals (a.u.))) for intensity models of each silicon line prior exclusion of non-sufficient models

3.5.3 Chapter summary

An overview of the best identification settings and the amount of identified/modelled lines is given in Table 8.
Concerning line identification, several difficulties were encountered. First, the lines in the database might not represent all lines present in the measured spectra apart from that only lines of the given material composition were modelled and not lines of hydrogen which were found as well. Furthermore, the amount of lines dropped significantly at identification and the identified lines did not always match the real peak shape due to a fixed peak half-width of the algorithm. This will result in simulated spectra with way fewer lines than measured ones and some artefacts. However, fixing this error would exceed the scope of this thesis.

Regarding line modelling, except for the dataset of borosilicate glass, way fewer line models were achieved for profile models than for simple intensity models. This reinforces the gap between present and simulated lines. The largest mismatch was presented in the dataset of aluminium alloy. However, the developed models were still used to simulate a complete spectrum in the next subsection.
Table 8 shows an overview of the optimised identification settings and the number of remaining models after each step as well as the suggested models for simulation.

Table 8: Overview of optimised line identification parameters and amount of identified/modelled lines for low-alloyed steel (SUS-1R), aluminium alloy (ERM-EB316) and borosilicate glass (NIST-1411)
* used for simulation in section 3.6 for spectrum simulation

Parameter	SUS-1R	ERM-EB316	NIST-1411
window width of the median filter (nm)	12	15	15
window width of the standard deviation filter (nm)	10	4	4
peak half width (nm)	0.07	0.07	0.07
line shift (nm)	-0.02	0.01	0.01
no. of lines in database	334	237	315
no. of identified lines	189	141	197
no. of fitted line profiles	171	120	187
no. of line profile models (LE and GD)	14	2	63
no. of line profile models (LE or GD)	84*	26*	169*
no. of intensity models (LE and GD)	185*	110*	114
no. of intensity models (LE or GD)	186	110	194*

3.6 Spectrum

After generating models for noise, baseline and signals according to Equation 19, all three models can be combined to predict a complete spectrum. The prediction workflow as well as some special features of it are discussed below.
Afterwards, the simulated spectra of each dataset are examined.

Simulation

For simulating noise and baseline, the intensity was predicted in dependency on gate delay and laser energy for each wavelength of the given range and summed up afterwards. The general simulation of the signal was based on an iteration over all available models and a summation of the predicted signals subsequently. The details differed between profile and intensity models. For the profile model, the corresponding A, σ and θ were predicted based on given gate delay and laser energy. These values were then subjected to Equation 6, whereas database reference values were used as λ_0. So, for the given wavelength range, the intensity was predicted for this line. In the case of intensity models, artificial line broadening was introduced to transform the simulated lines from simple straight signals to realistic line profiles. Therefore, the median σ value of the profile models was calculated and used as input for the Gaussian profile function (Equation 4). Again, the database reference value of the elemental line was used as λ_0.
After all an intensity value is achieved for each wavelength.

Exclusion of extreme prediction

Especially for profile models, lines with extremely high or low-intensity values were predicted occasionally. To exclude these lines, a recursive algorithm was developed.
In the first step, the local minima and maxima of the simulated spectrum were calculated. Then, local maxima value five times greater than the median local maxima value and local minima below zero were extracted. Afterwards, line models, whose λ_0 match the corresponding wavelengths of the local minima/maxima, were identified. These models were then excluded from the prediction in the second step.

The identification of extreme peaks was only performed once since it was observed that normal lines get identified as extreme lines if another recursion step takes place.

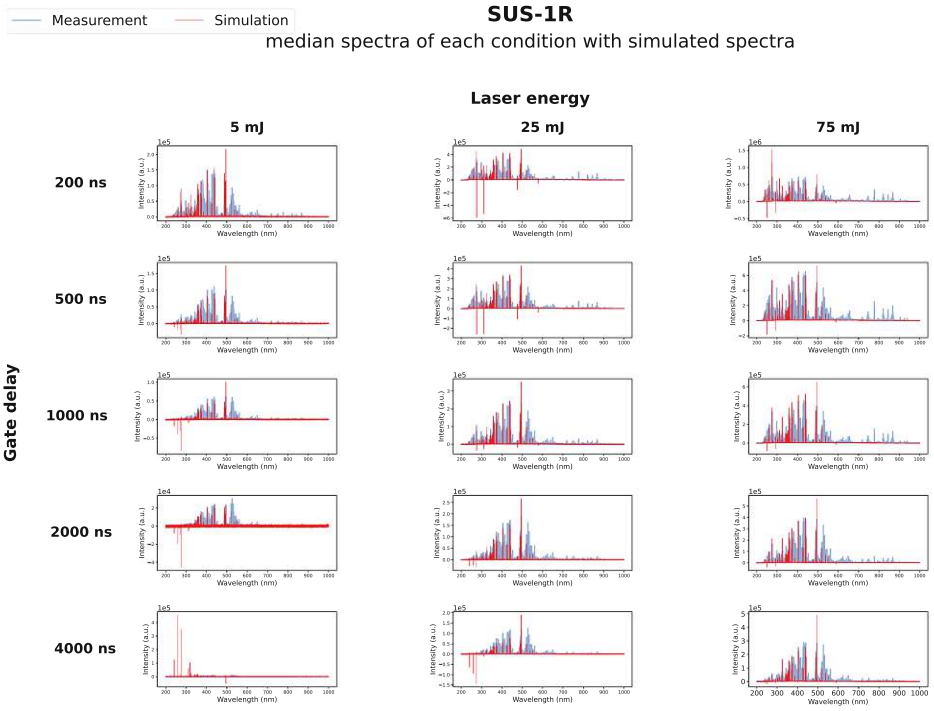
3.6.1 Simulation

In general, the simulated spectra can be distinguished between being based on modelled or artificial profiles and if the automated exclusion algorithm, as described in section 3.6, was applied or not. Moreover, for each dataset, the total spectrum and a detailed region (256.5 nm-262.5 nm) is displayed whereas the spectrum for the condition of maximum gate delay and minimum laser energy is additionally shown in a separate figure. The detailed region was used to analyse the profile fit.

3.6.1.1 Low-alloyed steel (SUS-1R)

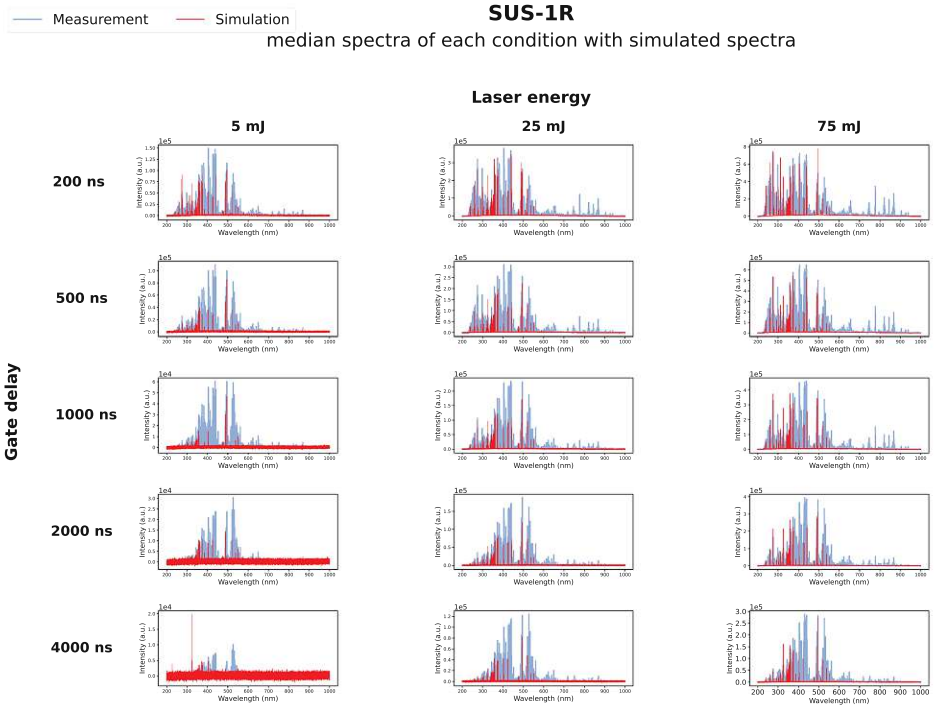
Modelled profiles

The comparison of Figure 47a and 47b or Figure 47c and 47d shows the effect of the automated exclusion algorithm. Obviously, extreme positive as well as negative lines get excluded successfully. For the detailed plots (Figure 47c, 47d) it can be seen that the predicted profiles and noise fits well to measured data. The baseline fits as well. However, as expected, a gap in the amount of measured and predicted lines is visible due to the discussed issues in section 3.5.3. Figure 48 shows enlarged plots of the condition with minimum gate delay and maximum laser energy.

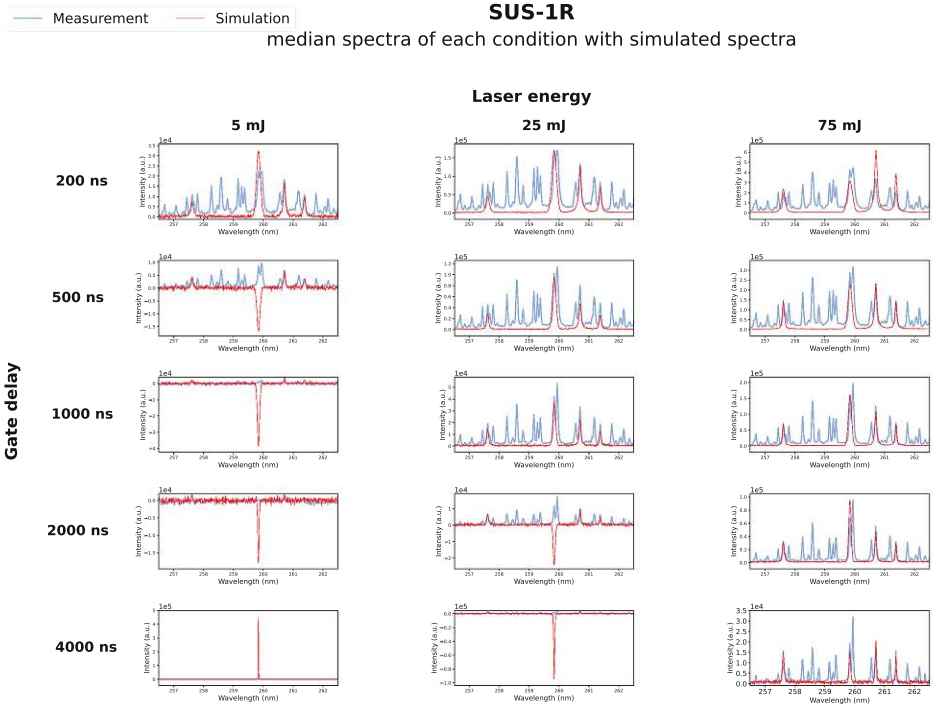


(a) total spectrum (auto-exclude=False)

Figure 47a: Low-alloyed steel (SUS-1R): simulation (red) of total spectrum (a, b) and detailed region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

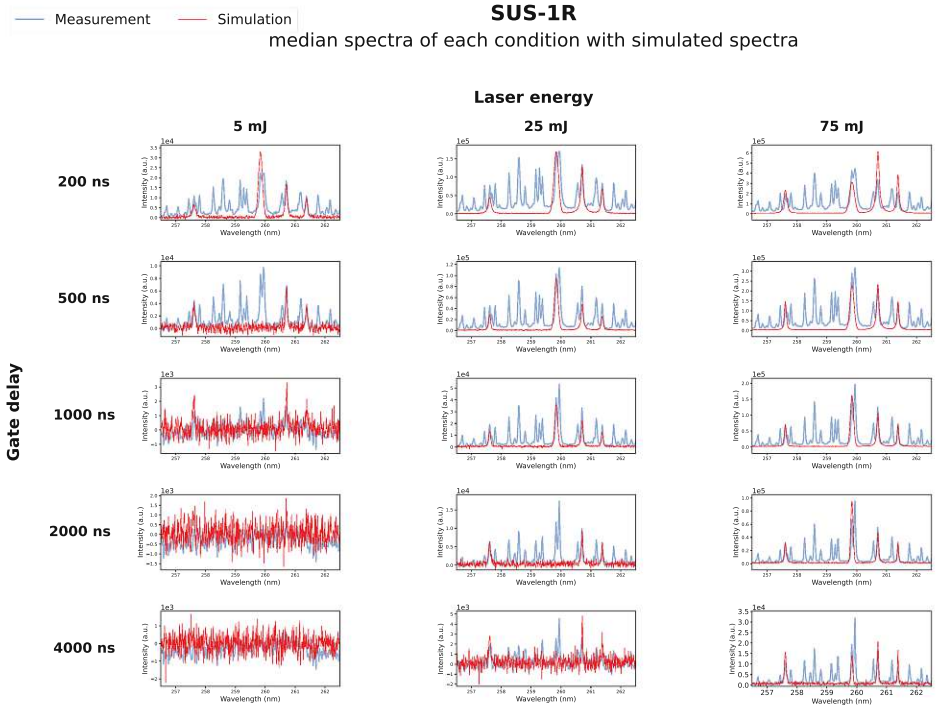


(b) total spectrum (auto-exclude=True)



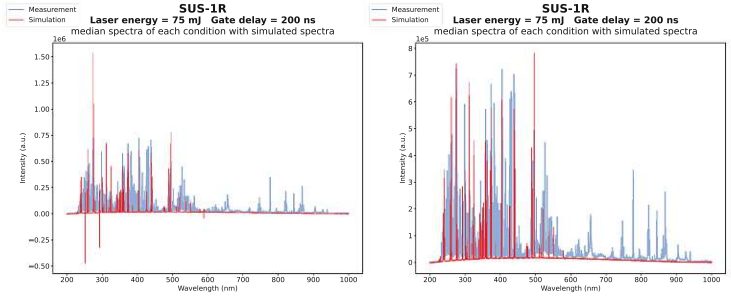
(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)

Figure 47b,c: Low-alloyed steel (SUS-1R): simulation (red) of total spectrum (a, b) and detailed region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



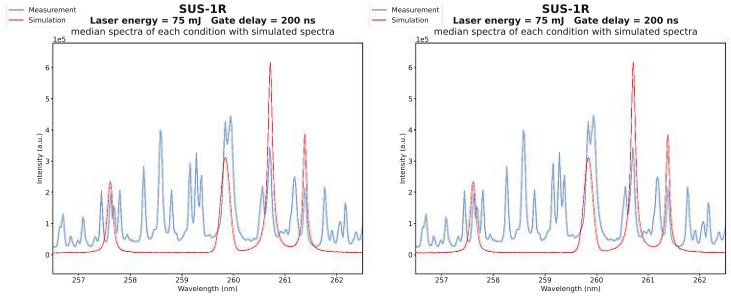
(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 47d: Low-alloyed steel (SUS-1R): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



(a) total spectrum (auto-exclude=False)

(b) total spectrum (auto-exclude=True)



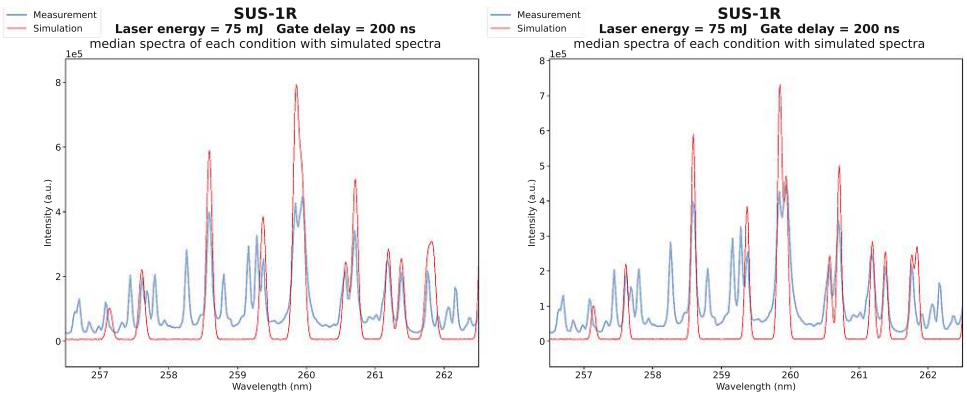
(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)

(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 48: Low-alloyed steel (SUS-1R): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) at 200 ns gate delay and 75 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

Artificial profiles

To simulate artificial profiles, the median σ of the profile models was calculated. Even though it was calculated as 0.04 nm, Figure 49a shows in comparison to Figure 49b that a value of 0.03 nm fits better. So, σ was set to 0.03.

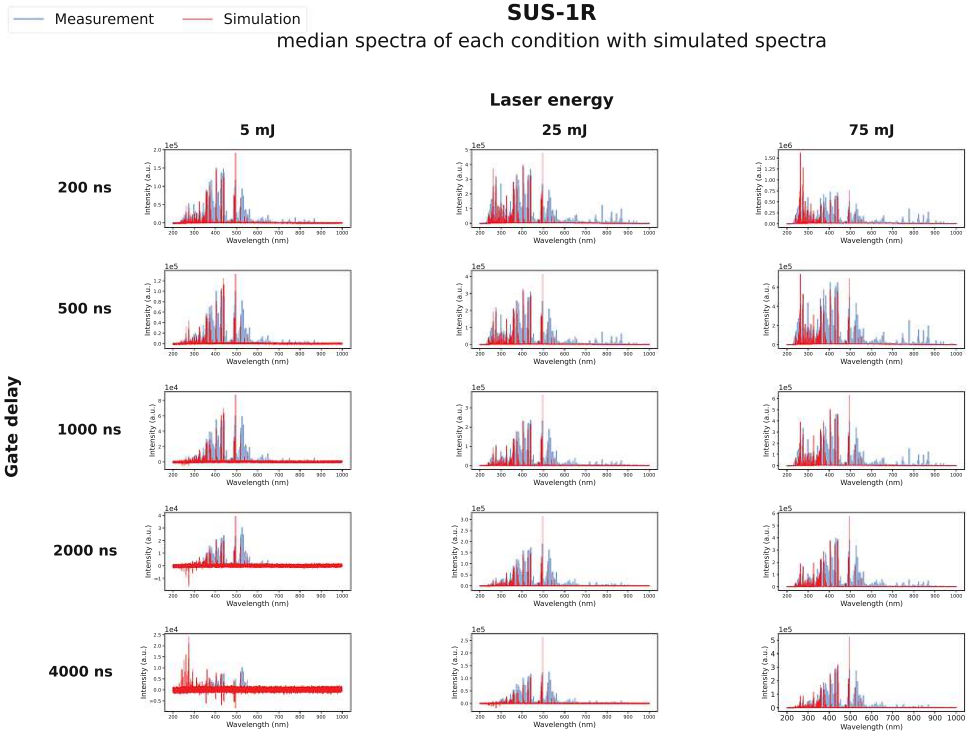


(a) σ=0.04 nm **(b)** σ=0.03 nm

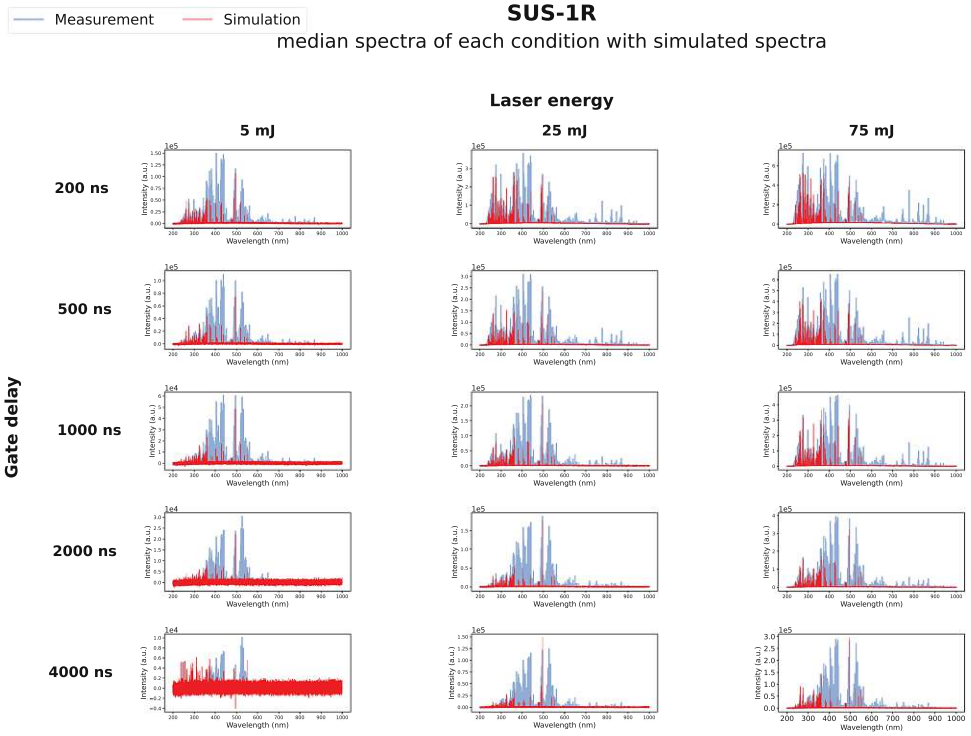
Figure 49: Low-alloyed steel (SUS-1R): simulation (red) of the detailed region based on artificial profiles with σ=0.04 nm (a) and 0.03 nm (b) and without automated exclusion of extreme predictions against median, measured spectrum (blue) at 200 ns gate delay and 75 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

Comparing Figure 50a to Figure 50b, the beneficial effect of the exclusion algorithm can be especially seen for the condition of 5 mJ laser energy and 4000 ns gate delay. But, one negative line still remains. Moreover, the decrease of normal-sized lines for the conditions of 1000-4000 ns GD and 25 mJ LE, emphasizes the conclusion that further tuning of the algorithm will be necessary in the future.
Regarding the spectra of the detailed region in Figure 50c and 50d, it can be seen that the artificial line broadening matches with measured data. Concerning the application of the automated exclusion algorithm, differences in line resolution can be found. So, the lines at 259.8 nm were better resolved without the exclusion algorithm. This finding is of special interest since no exclusion of extreme lines has taken place in this region. Algorithm tuning will be essential to solve this issue.
In general, more lines were visible in this spectra reflecting Table 8.
Figure 51 shows the magnified spectra of the condition of minimum gate delay and maximum laser energy.

Overall, the simulation with artificial profiles without the application of the automated exclusion algorithm showed the best results due to the increased amount of modelled lines compared to the profile models.

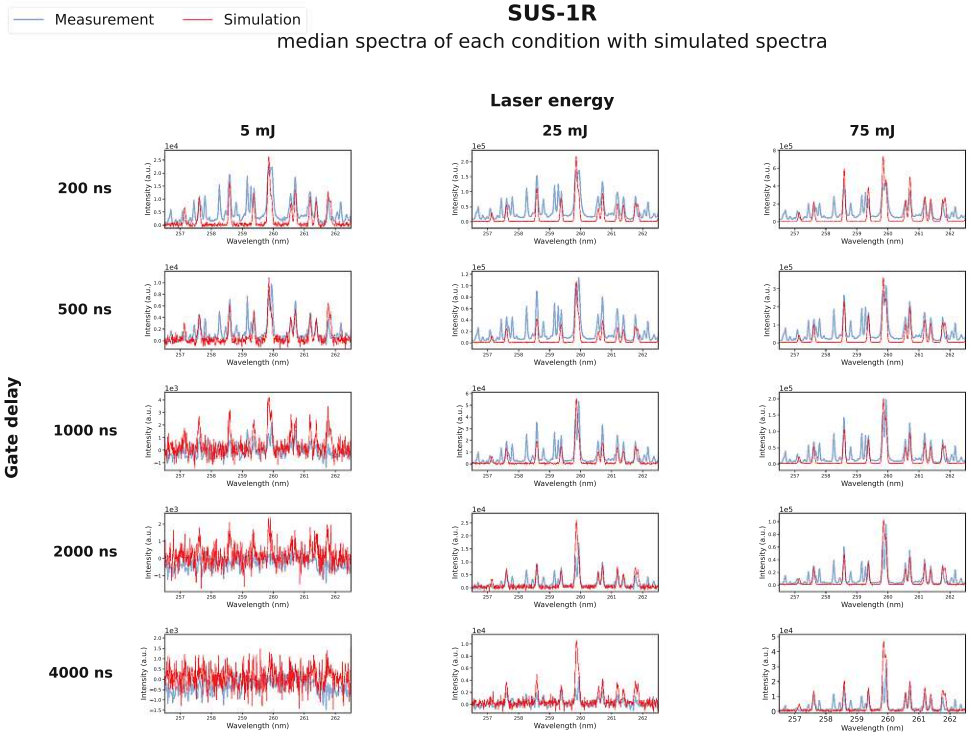


(a) total spectrum (auto-exclude=False)

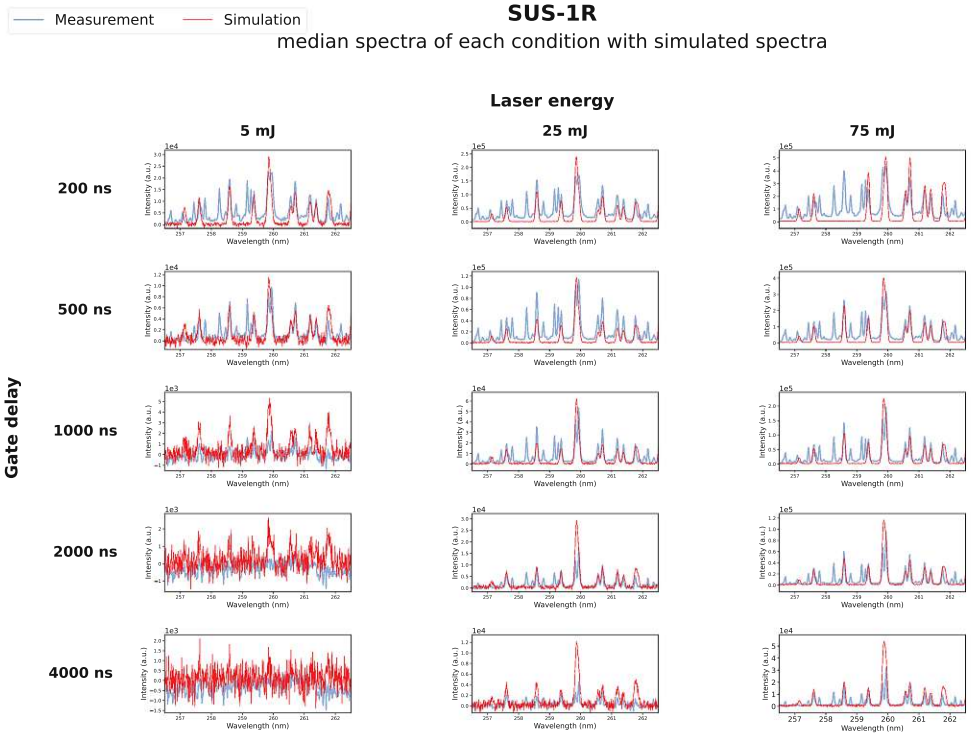


(b) total spectrum (auto-exclude=True)

Figure 50a,b: Low-alloyed steel (SUS-1R): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.03 nm) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

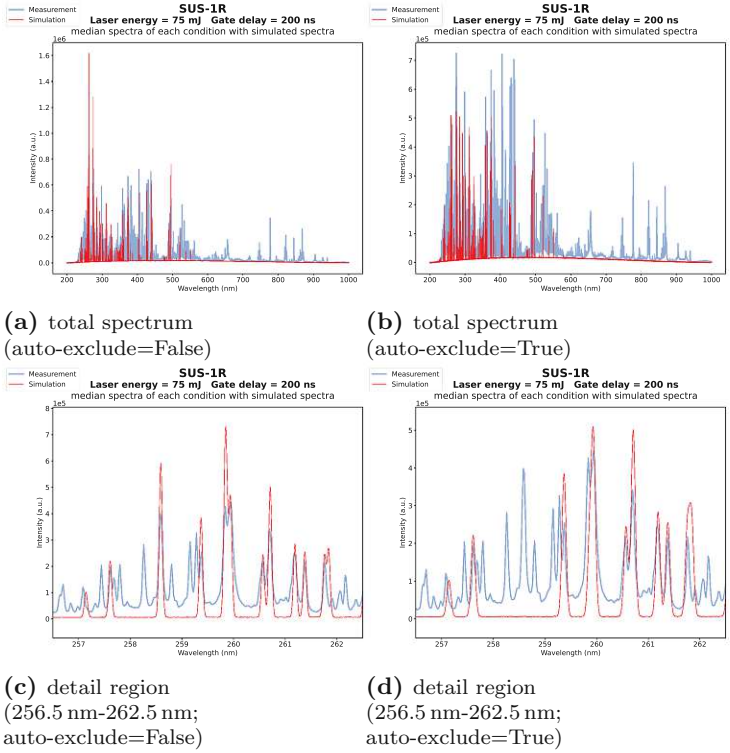


(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)



(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 50c,d: Low-alloyed steel (SUS-1R): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.03) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



(a) total spectrum (auto-exclude=False)

(b) total spectrum (auto-exclude=True)

(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)

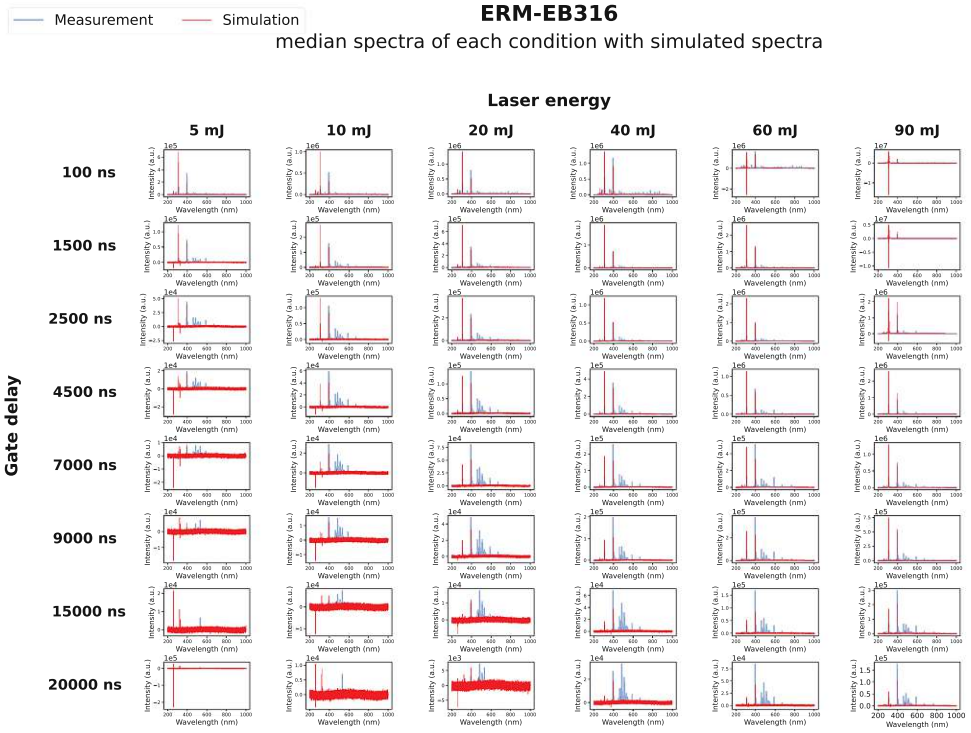
(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 51: Low-alloyed steel (SUS-1R): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.03 nm) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) at 200 ns gate delay and 75 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

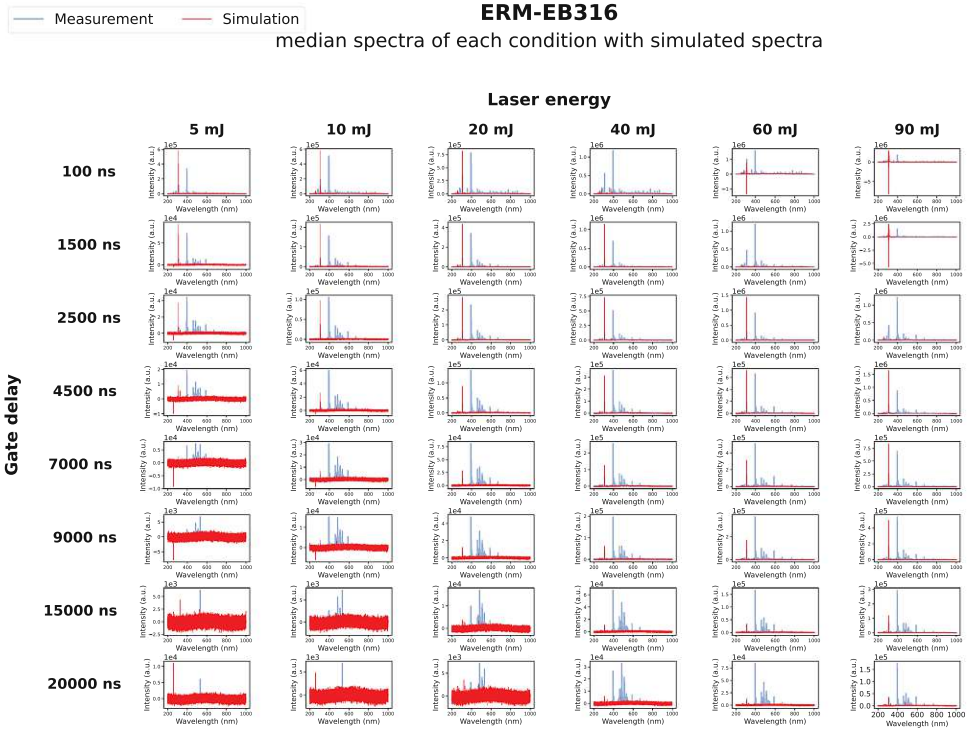
3.6.1.2 Aluminium alloy (ERM-EB316)

Modelled Profiles

As a result of the data given in Table 8, the simulated spectra (Figure 52a,b lack significantly in the number of lines. Again, it is seen that the automated exclusion algorithm does not apply to all wrong predicted peaks. This can be shown especially for the condition at 100 ns gate delay and 90 mJ laser energy. Enlarged versions of this condition are displayed in Figure 53.

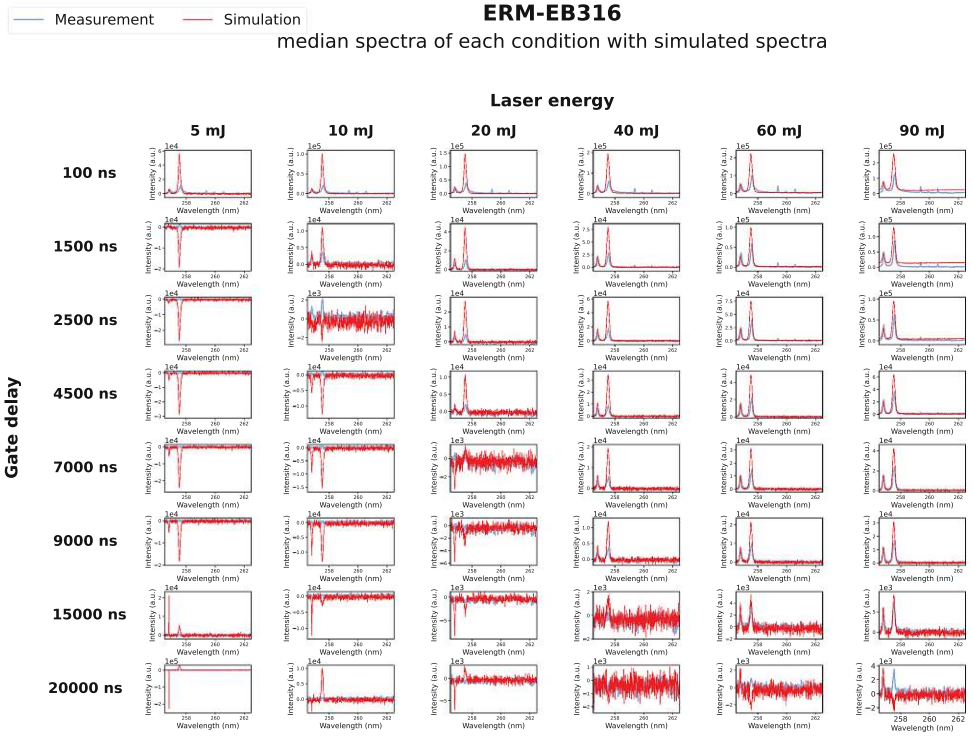


(a) total spectrum (auto-exclude=False)

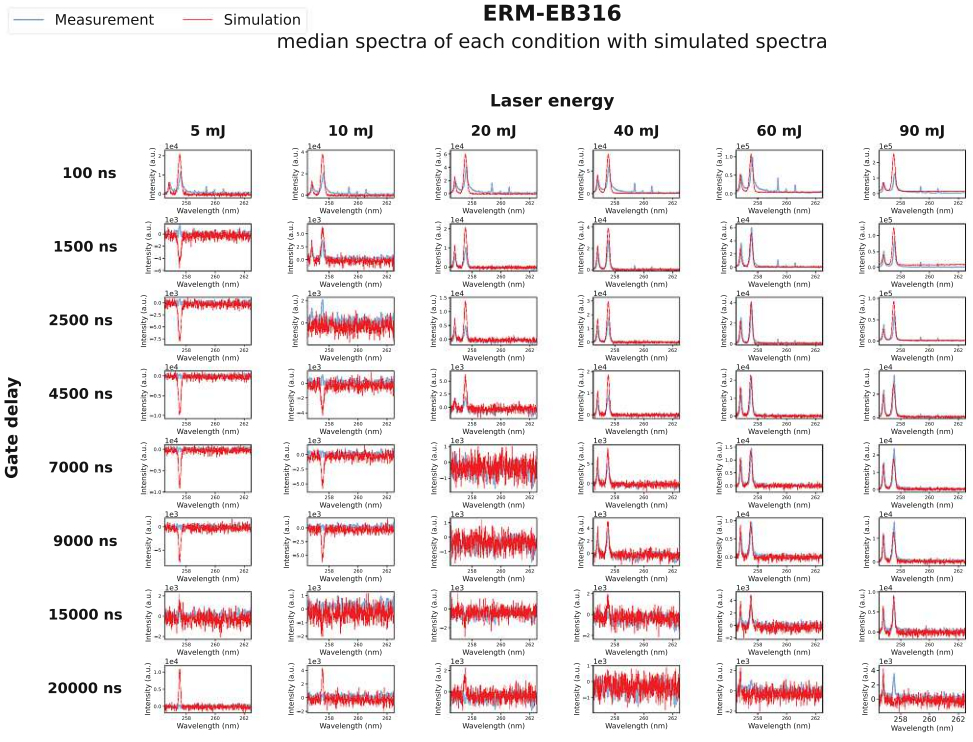


(b) total spectrum (auto-exclude=True)

Figure 52a,b: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

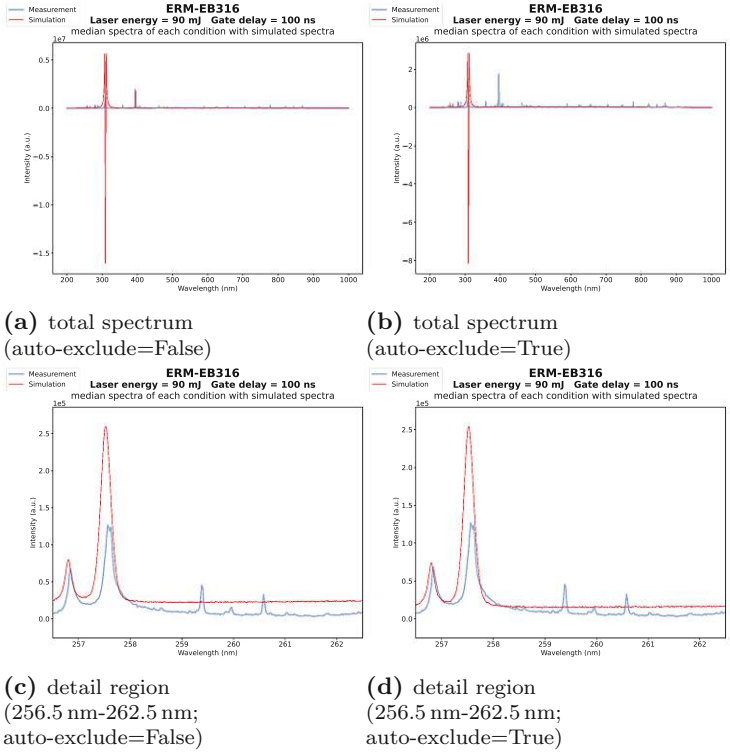


(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)



(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 52c,d: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



(a) total spectrum (auto-exclude=False)

(b) total spectrum (auto-exclude=True)

(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)

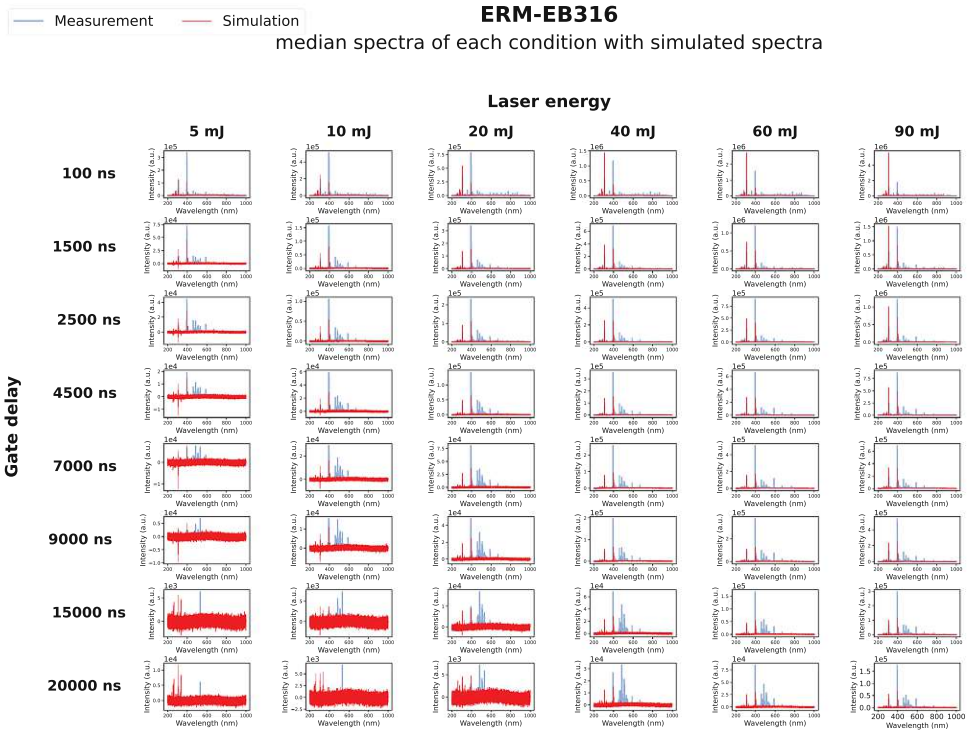
(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 53: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on modelled profiles with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) at 100 ns gate delay and 90 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

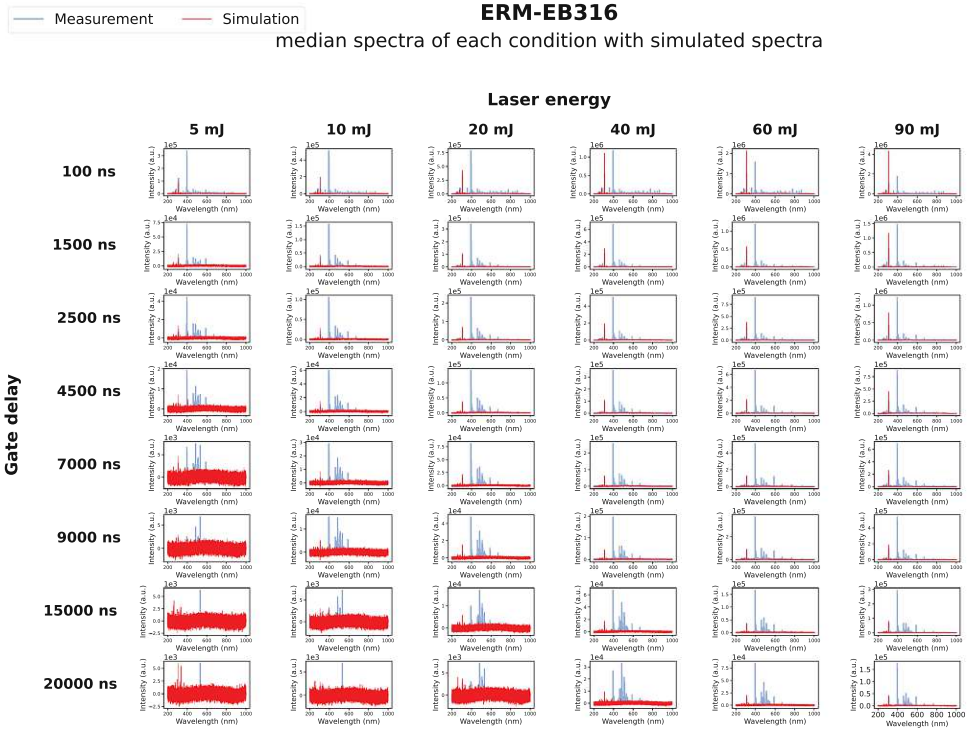
Artificial profiles

The median σ of the profile models was calculated as 0.05 nm and resulted in good artificial profiles (Figure 55c and 55d). Concerning the total spectra in Figure 54a and 54b, way fewer lines than measured were predicted again. Nevertheless more lines than for the profile models in Figure 52a,b were present. Once more, the automated exclusion algorithm was not able to fix all lines. This problem is better visible in Figure 55a and Figure 55b for the line at 310 nm.

Anyhow, simulations with artificial profiles without the application of the exclusion algorithm resulted in the best simulations for this material due to the larger number of lines compared to profile models and the possibility of neglecting the exclusion algorithm due to its errors.

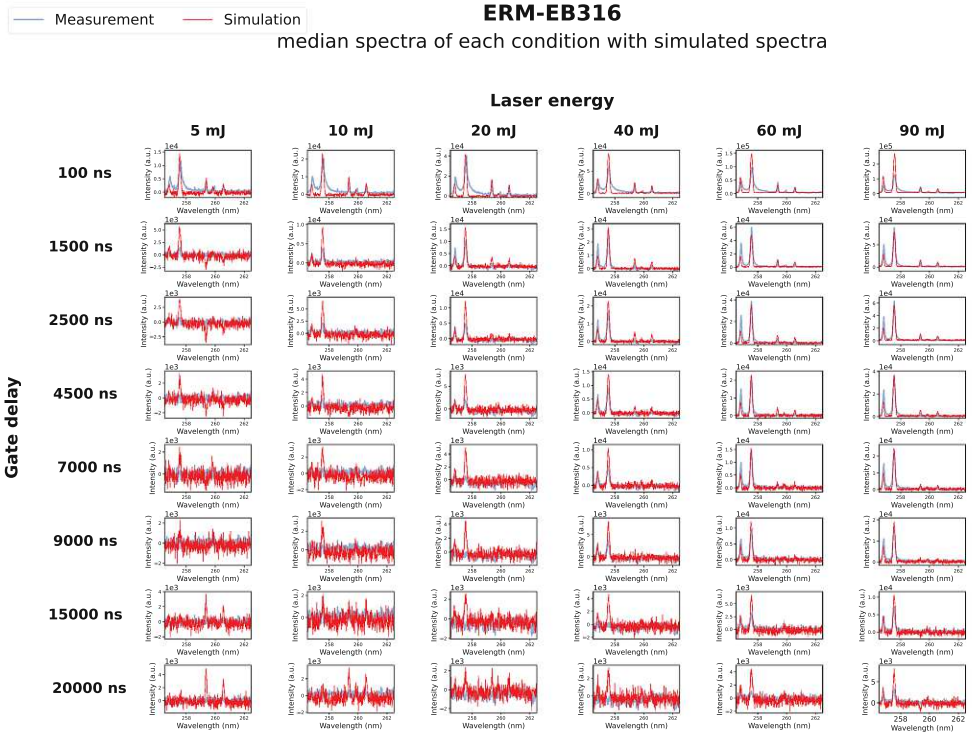


(a) total spectrum (auto-exclude=False)

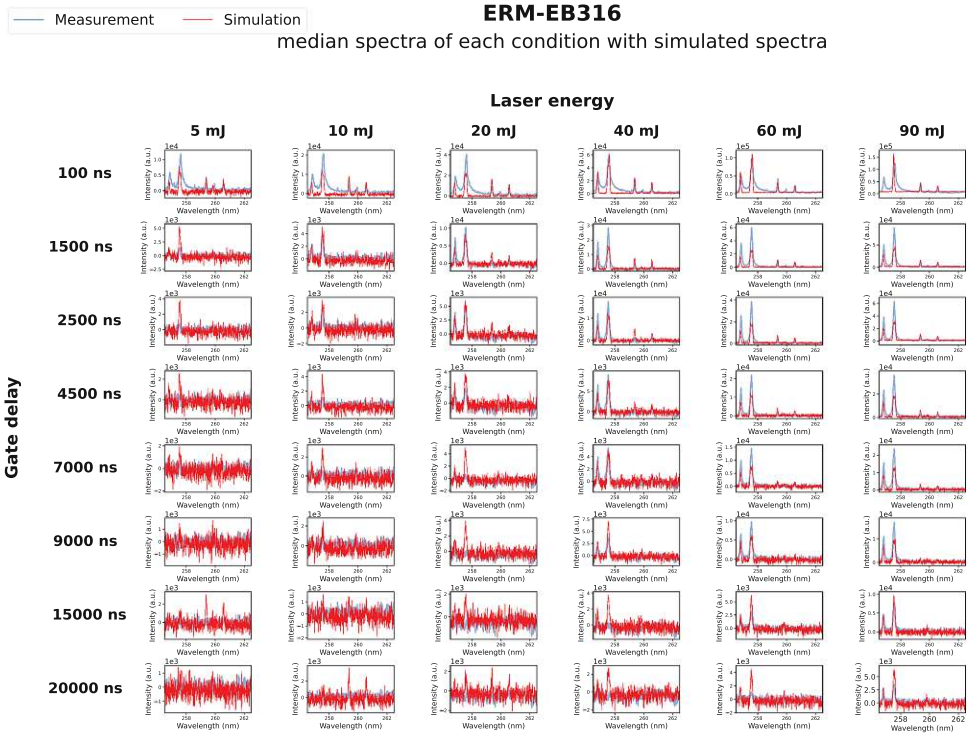


(b) total spectrum (auto-exclude=True)

Figure 54a,b: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.05 nm) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

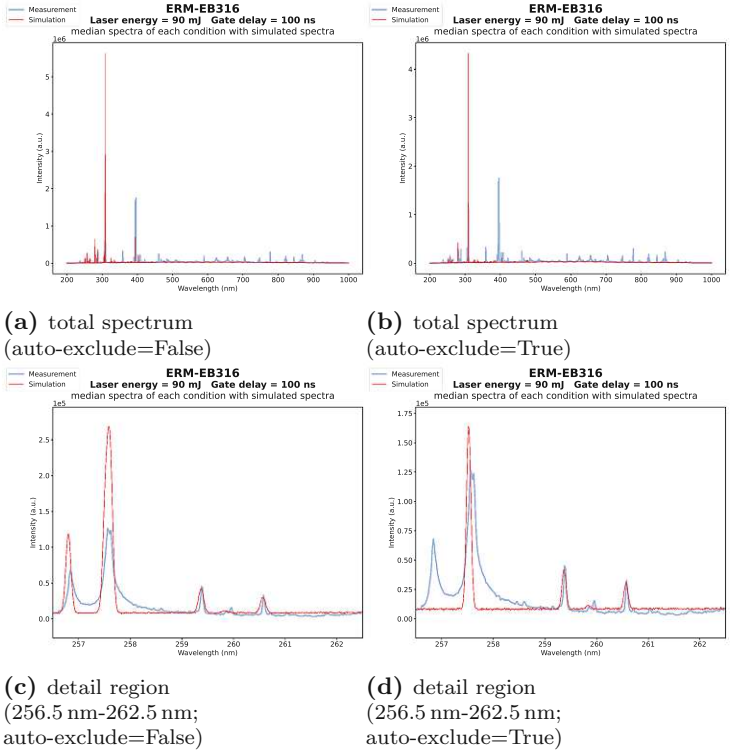


(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)



(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 54c,d: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.05 nm) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



(a) total spectrum (auto-exclude=False)

(b) total spectrum (auto-exclude=True)

(c) detail region (256.5 nm-262.5 nm; auto-exclude=False)

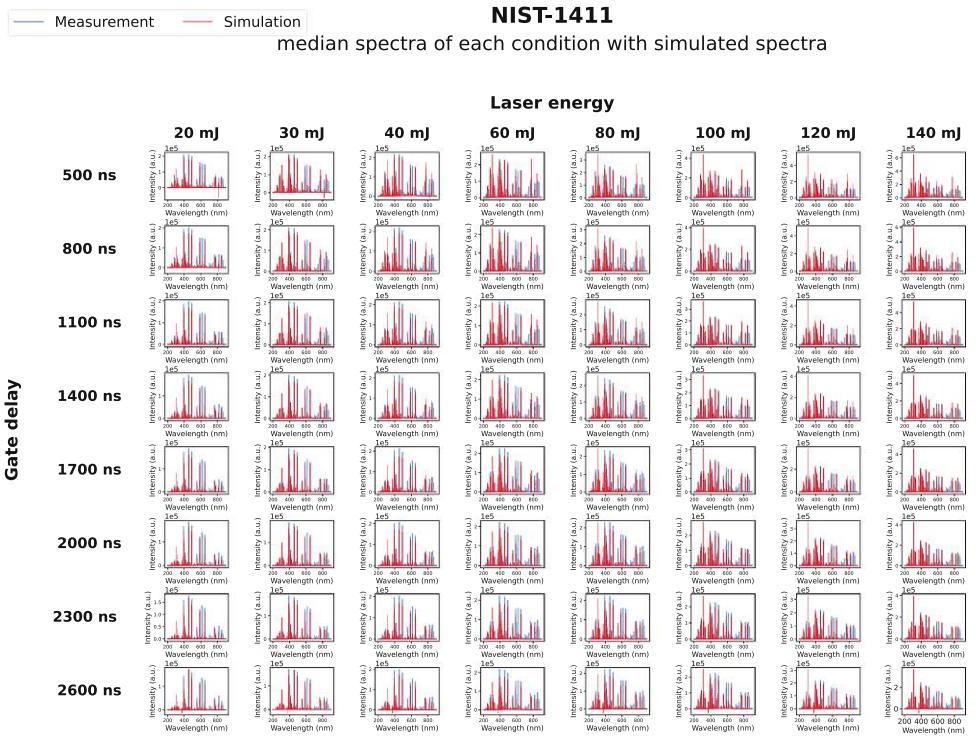
(d) detail region (256.5 nm-262.5 nm; auto-exclude=True)

Figure 55: Aluminium alloy (ERM-EB316): simulation (red) of the total spectrum (a, b) and detailed simulated region (c, d) based on artificial profiles (σ=0.05 nm) with (a, c) and without (b, d) automated exclusion of extreme predictions against median, measured spectrum (blue) at 100 ns gate delay and 90 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

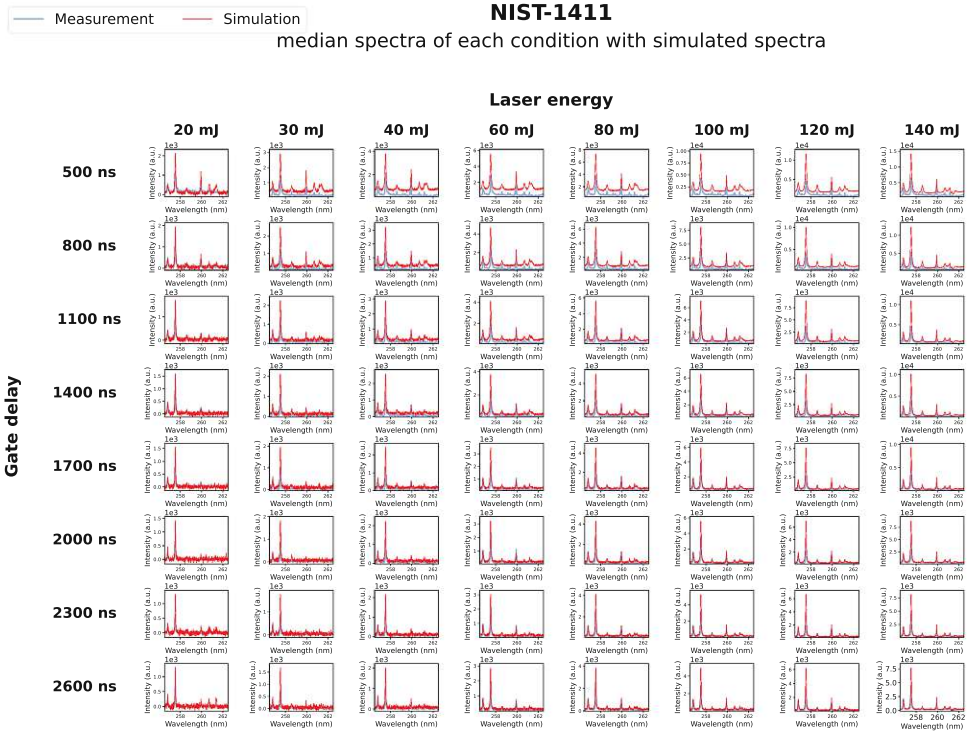
3.6.1.3 Borosilicate glass (NIST-1411)

Modelled profiles

For this dataset, proportionally many lines could be simulated due to the relatively high yield of successful models (see Table 8). Since no extremely misfitted lines were present, the automated exclusion algorithm was not applied. The predicted lines fitted in terms of intensity and line profile (Figure 56). The baseline offset in Figure 57b is not of importance, since it is constant across the wavelength range.

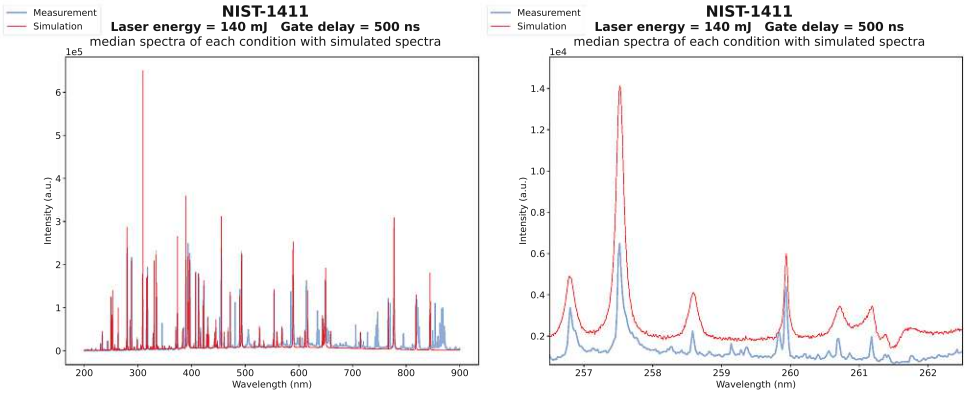


(a) total spectrum (auto-exclude=False)



(b) detail region (256.5 nm-262.5 nm; auto-exclude=False)

Figure 56: Borosilicate glass (NIST-1411): simulation (red) of the total spectrum (a) and detailed simulated region (b) based on modelled profiles without automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



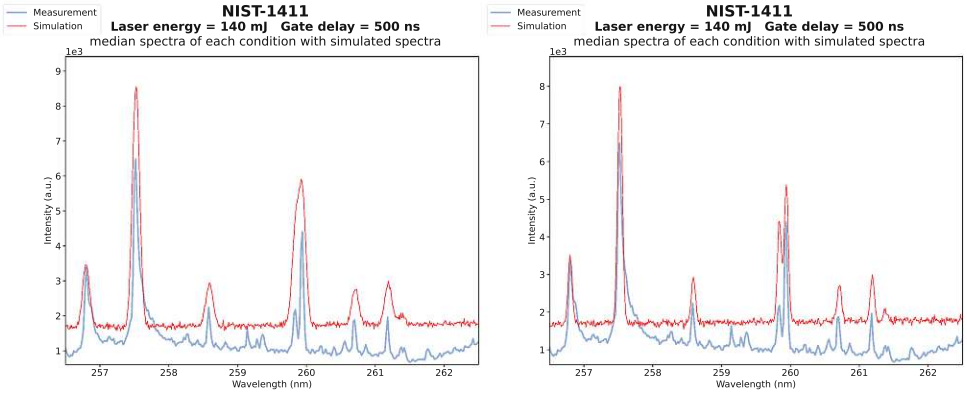
(a) total spectrum (auto-exclude=False) **(b)** detail region (256.5 nm-262.5 nm; auto-exclude=False)

Figure 57: Borosilicate glass (NIST-1411): simulation (red) of the total spectrum (a) and detailed simulated region (b) based on modelled profiles without automated exclusion of extreme predictions against median, measured spectrum (blue) at 500 ns gate delay and 140 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

Artificial profiles

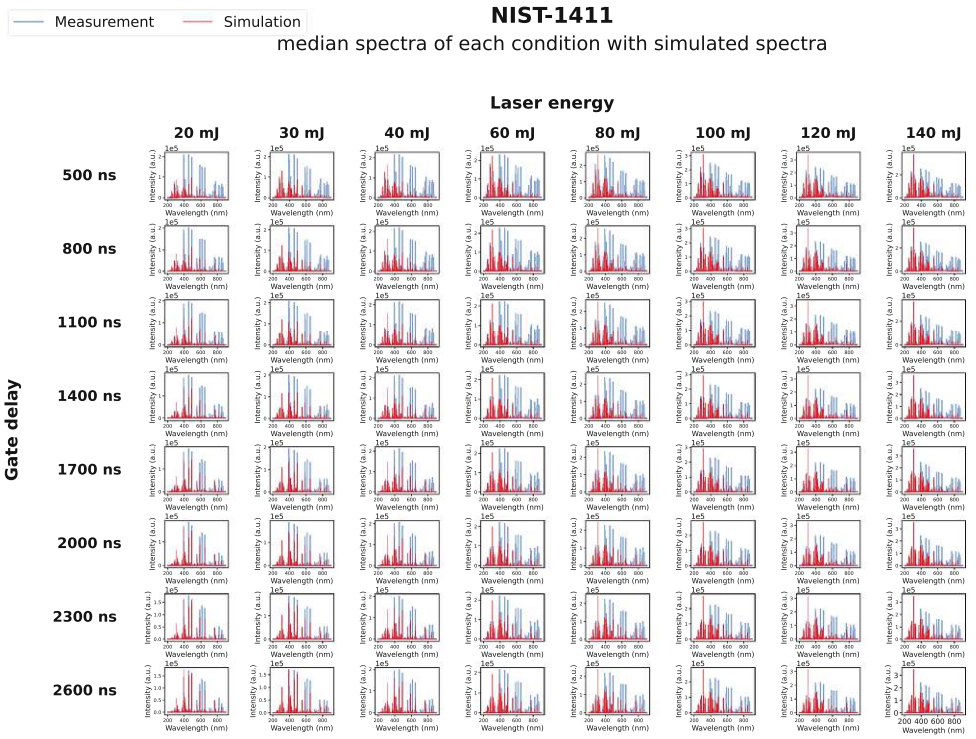
The median σ was calculated as 0.05 nm. But predicted lines based on this value were too broad (Figure 58a). So, the final σ value was set as 0.03 nm (Figure 58b). Considering the dataset plots (Figure 59), no misfitted lines could be identified, again. Consequently, the automated exclusion algorithm was not applied.
Overall, the predicted lines were less intense than the measured ones but the profile shape fitted well. In the magnified plots of the condition of minimum gate delay and maximum laser energy in Figure 60, a slight offset is visible. As for the profile models, this offset can be accepted.

Given the fact that slightly more lines could be simulated by artificial profiles, this simulation is superior to the modelled profiles.

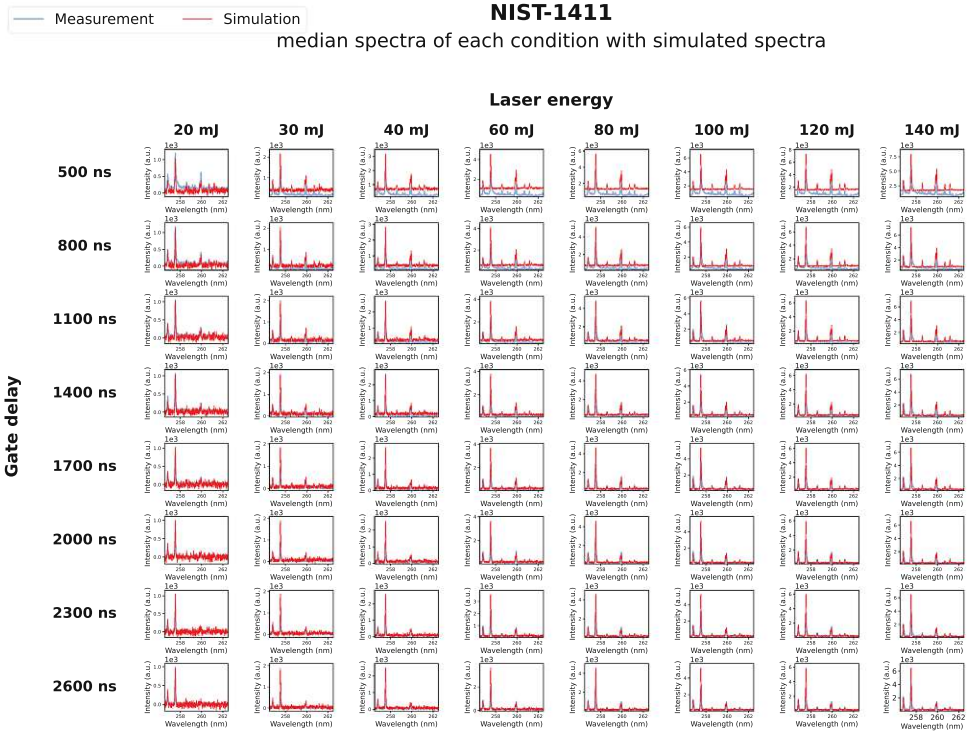


(a) σ=0.05 nm **(b)** σ=0.03 nm

Figure 58: Borosilicate glass (NIST-1411): simulation (red) of the detailed region based on artificial profiles with σ=0.05 nm (a) and 0.03 nm (b) and without automated exclusion of extreme predictions against median, measured spectrum (blue) at 500 ns gate delay and 140 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

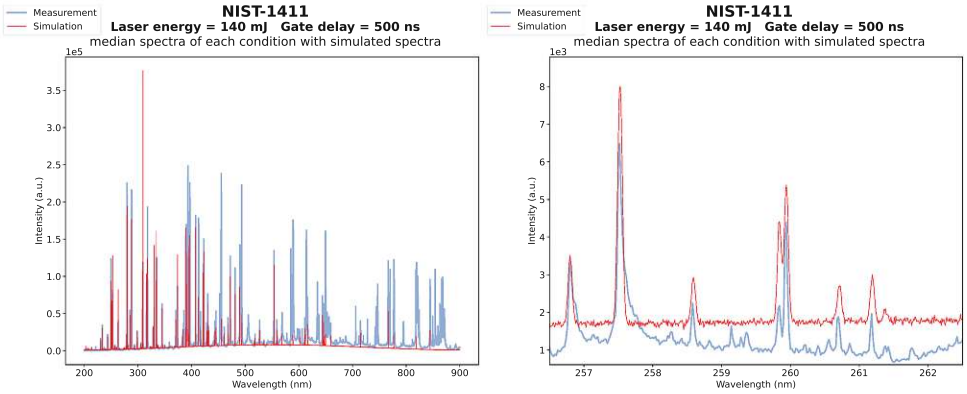


(a) total spectrum (auto-exclude=False)



(b) detail region (256.5 nm-262.5 nm; auto-exclude=False)

Figure 59: Borosilicate glass (NIST-1411): simulation (red) of the total spectrum (a) and detailed simulated region (b) based on artificial profiles (σ=0.03 nm) without automated exclusion of extreme predictions against median, measured spectrum (blue) for all conditions of the dataset; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)



(a) total spectrum (auto-exclude=False) **(b)** detail region (256.5 nm-262.5 nm; auto-exclude=False)

Figure 60: Borosilicate glass (NIST-1411): simulation (red) of the total spectrum (a) and detailed simulated region (b) based on artificial profiles (σ=0.03 nm) without automated exclusion of extreme predictions against median, measured spectrum (blue) at 500 ns gate delay and 140 mJ laser energy; Plots: $\tilde{I}_c$ (a.u.) vs. λ (nm)

3.6.2 Chapter summary

Considering the simulated spectra of all datasets, common successes and shortcomings could be identified. First, the simulated noise and baseline blended in satisfactorily. The predicted line profile independent from modelling or artificial generation, fitted the measured data quite well. In terms of intensity, the results were mixed depending on the dataset and modelling method, but overall the shape of the spectra could be predicted. Furthermore, the automated exclusion algorithm performed well for the majority of data. However, the failures of the algorithm also lead to major setbacks in modelling since misfitted lines can distort the entire spectrum. By now, it should be further investigated why some obviously mispredicted lines did not get excluded and why the algorithm sometimes changes the line resolution - as observed for the low-alloyed steel dataset. Additionally, the algorithm should be adapted to avoid the exclusion of lines, if all lines fit well.
An overview of the settings for each model type and material as well as the overall best models is given in Table 9. Comparing the datasets to each other, the total best simulation was achieved for borosilicate glass. Here, the results for modelled and artificial profiles were almost equal. Yet, slightly more lines could be modelled for artificial profiles (Figure 57b versus 60b). The worst results were obtained for the dataset of the aluminium alloy. Since this dataset had the second most measurement conditions available and thereby the second most amount of data to build the models, the size of provided datasets cannot explain the difference in final simulation quality. One reason to explain the poor performance of aluminium alloy, could be difficulties in line identification if Figure 36a is compared to Figure 35a or 37a. This falls back to the previously discussed issues of the applied line identification algorithm in section 3.5.3.

Table 9: Overview of the best simulation settings for each model type (modelled or artificial profiles) for low-alloyed steel (SUS-1R), aluminium alloy (ERM-EB316) and borosilicate glass (NIST-1411)
* selected as best simulation model

Material	modelled profiles	artificial profiles
SUS-1R	with exclusion algorithm	without exclusion algorithm (σ=0.03)*
ERM-EB316	with exclusion algorithm	without exclusion algorithm (σ=0.05)*
NIST-141	without exclusion algorithm	without exclusion algorithm (σ=0.03)*

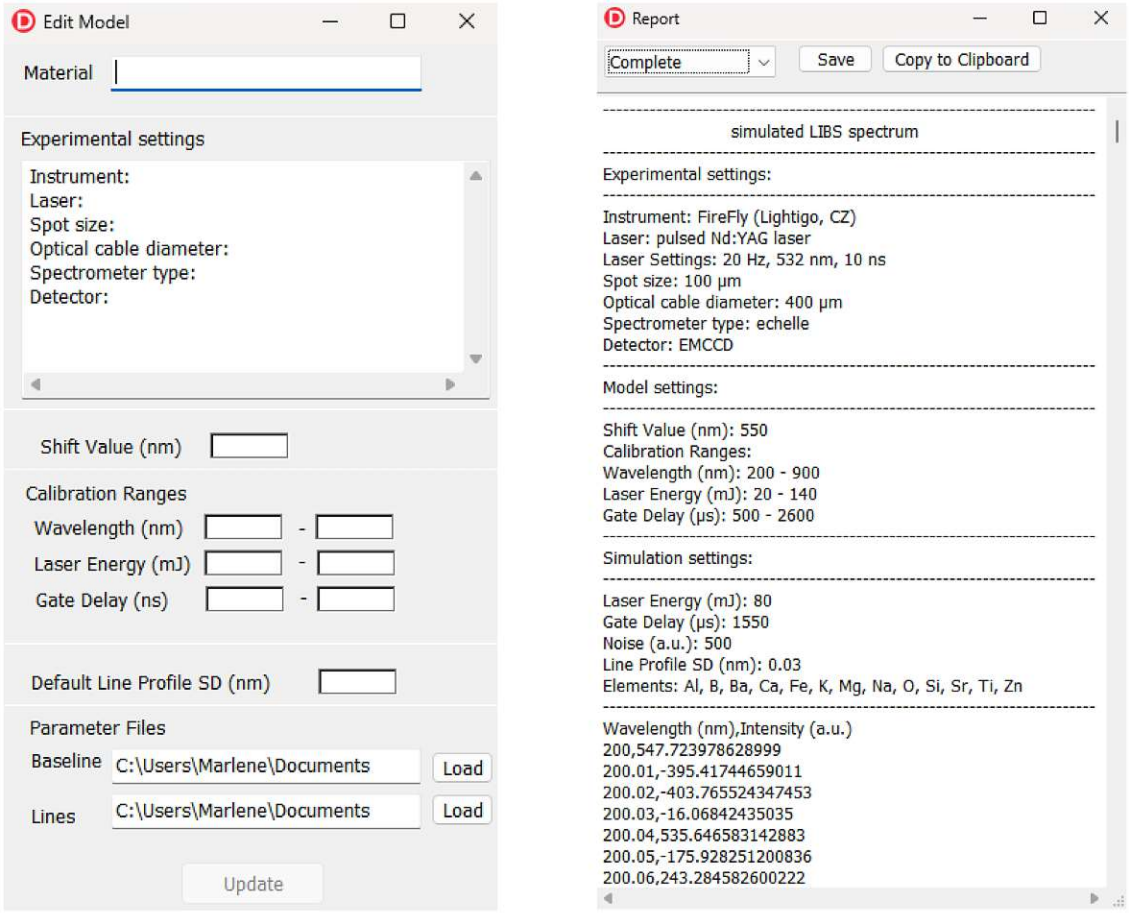
The automated exclusion algorithm worked well for the majority of extreme lines.

3.7 User Interface

To explore the generated models, a custom user interface was developed in Delphi/Object Pascal with the Empacadero integrated development environment (Delphi 10.4). The user interface is available for download at `https://zenodo.org/records/10557230`.
In the following, an overview of the input data to run the program as well as the user interface will be given.

3.7.1 Input data

To simulate a LIBS spectrum of a material, the information requested as in Figure 61a must be initially given to transfer the model, developed in Python, to the software. The shift value is usually the mean value of the wavelength calibration range as discussed in section 3.4.1 and Equation 21. The default line profile standard deviation is the σ value as given in Table 9. Since the models with artificial line profiles were considered as the best models for each dataset, these model types were used for simulating the LIBS spectra in the custom-built software. Therefore, text files with all parameters of the baseline and line models had to be exported from the Python models (automatically performed at model fit) and needed to be loaded as input files.



(a) window to edit/create new model **(b)** window to export report

Figure 61: Screenshots of custom-built user interface

3.7.2 Features

In Figure 62, the main window with the available features is displayed. For the simulation, it is possible to adjust the gate delay, the laser energy, the noise level, the width of the Gaussian line profile and to select the included elements. For the visual display, zoom buttons are available and there is an option for auto-scaling the spectra. Additionally, negatively predicted lines can be excluded. Furthermore, a report can be generated. Figure 61b shows the report window including the option to export the complete report or only the settings or the spectrum values.

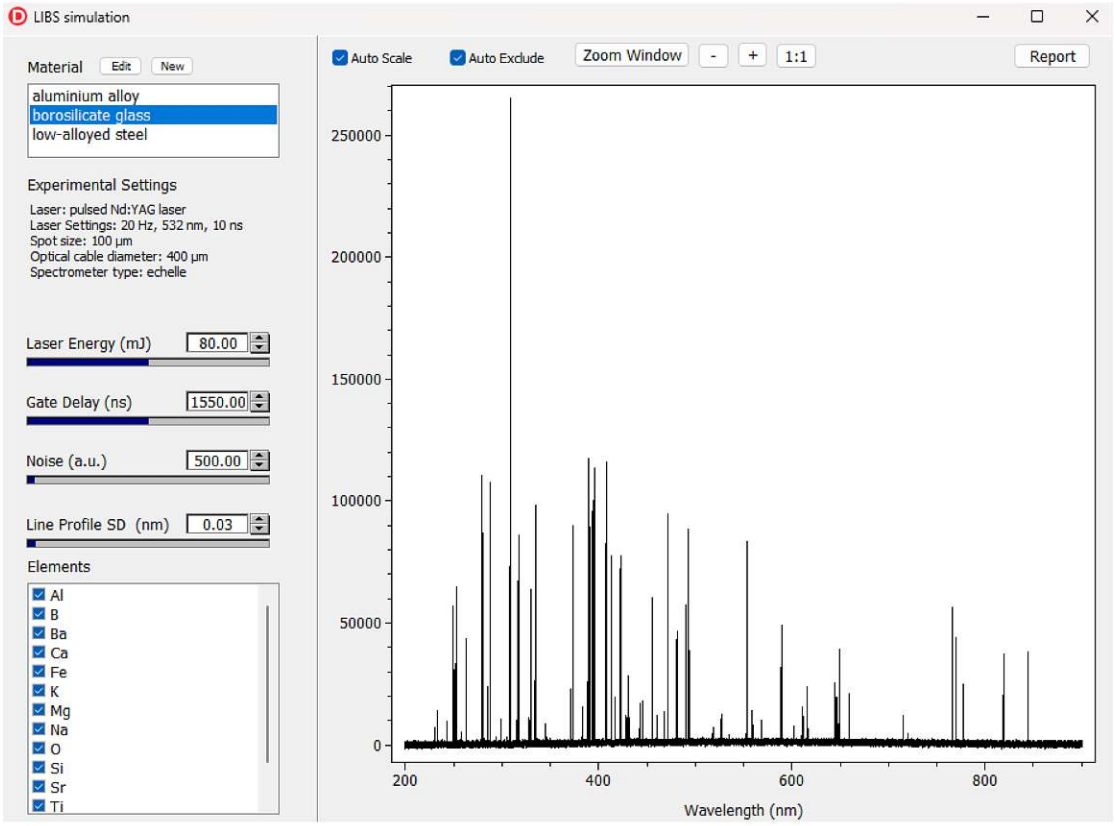


Figure 62: Screenshot of custom built user interface: main window

3.7.3 Chapter summary

To sum up, the user interface is capable of illustrating the influence of gate delay and laser energy as well as the presence of elements onto the LIBS spectra of different materials. However, some additional features to enhance the exploration experience like an overlay of specific elemental lines should be added in the future. Furthermore, some layout details can be improved like centring the data values in the report window (Figure 61b).

4 Conclusion

The aim of this thesis was to develop empirical models of LIBS spectra to enhance the handling of real-world variations opposed to conventional theoretical models relying on physicochemical phenoma. In pursuit of this objective, polynomials were chosen as the modelling framework in this study.

The general approach to model each of the three spectral constituents (the noise, the baseline, and the emission lines) separately prior to merging all together resulted in models for each provided dataset (low-alloyed steel, aluminium alloy, borosilicate glass). In the following, the general results for all three spectral components as well as the merged model are discussed whereas more detailed explanations are provided within the matching chapter summaries.

Despite the character of the noise showed to be heteroscedastic, normally distributed random errors representing homoscedastic noise were used to reduce the modelling effort. As a simulating parameter, the standard deviation of the Gaussian normal distribution in absolute intensity values was employed. For all three materials, the simulated noise blended in well.

For baseline simulation, polynomial baseline regression of the measured data was applied beforehand to estimate the underlying baseline polynomials. The polynomial degree was limited to five. The analysis revealed notable variations in the fundamental baseline polynomials among the three materials. Nevertheless, individual baseline fit for all conditions of each material could be achieved, utilizing at least a few baselines. The models derived from baseline regression data demonstrated effective prediction of the baselines. Some baselines exhibited a constant offset across the wavelength range. However, this was considered acceptable, as such offset only impacted the intercept and not the overall shape of the spectrum, ensuring that the quality of the simulation was not compromised.

Prior to the modelling of the signal, the emission lines of each material were identified, considering its elemental composition. However, the algorithm employed for line identification suffered several shortcomings at that stage, such as non-dynamic peak half-width. Consequently, the number of successfully identified lines deviated significantly from the theoretically available lines in the database. As addressing these issues was beyond the scope of this thesis, the discrepancies in the line identification were acknowledged and resulted in simplifications of the final models.
Two approaches were applied for modelling. In the first approach, a Pseudo-Voigt profile was fitted for each identified line and separate polynomial models for σ, A and θ in dependency of gate delay and laser energy were developed. The second approach involved was based on modelling of the peak intensity solely. In both cases, the maximum polynomial degree was limited to two. Moreover, the number of available line models dropped by the exclusion of non-sufficient models. The yields were notably lower for the Pseudo-Voigt line profile models compared to simple intensity line models. Using both approaches, the highest yield of adequate lines was consistently observed for borosilicate glass, highlighting its suitability for the proposed methodology. In contrast, the yield was considerably lower for aluminium alloy, emphasising material-specific variations in the effectiveness of the modelling techniques.

After establishing models for the individual spectral components, the final intensities were determined by their summation. For the final simulation, the line intensity models were extended with artificial line broadening based on a Gaussian profile. As λ_0, the database reference value was used and σ was set to match measured data. Concerning the simulated line profile, no difference between modelled and artificial profiles could be found. Despite the quality of predicted intensity varied throughout the datasets and modelling methods, the overall shape of the spectrum could be simulated irrespective of the applied line model. The application of a custom automated line exclusion algorithm to eliminate lines with predicted negative intensity or immoderate positive values was specifically useful for simulations based on profile models. Nevertheless, the algorithm still requires some adaptation since not all extreme lines were detected. Overall, the simulations based on artificial profiles without the application of the automated exclusion algorithm were of higher quality due to their higher amount of simulated lines.

Comparing the final simulation results of all three materials, the best performance was obtained for borosilicate glass and low-alloyed steel. The worst quality was achieved for aluminium alloy. Reasons for this might lie in the major difficulties of the line identification algorithm for aluminium alloy. The amount of provided data is not considered as a reason since aluminium alloy contained the second most measurements after borosilicate glass.

To conclude, a novel model approach for LIBS spectra was developed. For visualisation, a custom user interface was provided. Hence, the models can already be used to explore the rough impacts of gate delay and laser energy on LIBS spectra for the material classes of low-alloyed steel, aluminium alloy and borosilicate glass.

5 Outlook

The development of polynomial models for LIBS spectra has established a foundation for examining the impact of alterations in laser energy and gate delay on the characteristics of noise, baseline, and emission lines across various materials. Nonetheless, these models exhibit certain limitations that warrant future refinement. Specifically, improvements are needed in the line identification algorithm, as the allocated emission regions form the basis for constructing line models. A proposed enhancement involves the implementation of a dynamic peak half-width to better accommodate changes in peak widths within the spectrum. Furthermore, additional features in the user interface, such as an overlying mask highlighting elemental lines, should be incorporated to facilitate easier exploration of spectra.

Although the empirical approach of simulating LIBS spectra offers advantages in capturing real-world complexities and providing a practical understanding of laser-material interactions, it comes with significant limitations. These include the challenge of accurately representing non-linear relationships and intricate spectral patterns in complex scenarios. Therefore, exploring alternative strategies beyond physical and empirical simulation could offer a more comprehensive and adaptable solution to overcome these limitations.

One such promising avenue is the application of neural network models in LIBS spectra simulation, which merits in-depth exploration. These models demonstrate a notable capacity to learn intricate patterns and adapt to complex relationships in data, making them particularly well-suited for the non-linear and dynamic nature of LIBS spectra. However, it's important to acknowledge potential limitations, such as the need for substantial training data, computational resources, and challenges in interpretability. The success of neural networks often relies on large datasets for training, which may be challenging to obtain for certain LIBS applications. Moreover, the interpretability of neural network models can be challenging, emphasising the need to strike a balance between model complexity and transparency for practical real-life applications. The exploration of neural network models should carefully consider these requirements and potential limitations to ensure their effectiveness and feasibility in LIBS spectra simulation in real-world scenarios.

List of Figures

List of Tables

References

[1] Zuzana Gajarska et al. "Identification of 20 polymer types by means of laser-induced breakdown spectroscopy (LIBS) and chemometrics". In: *Analytical and Bioanalytical Chemistry* 413.26 (2021), pp. 6581–6594.

[2] Márcio José Coelho Pontes et al. "Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain". In: *Analytica Chimica Acta* 642.1-2 (2009), pp. 12–18.

[3] Zengqi Yue et al. "Machine learning-based LIBS spectrum analysis of human blood plasma allows ovarian cancer diagnosis". In: *Biomedical optics express* 12.5 (2021), pp. 2559–2574.

[4] Celio Pasquini et al. "Laser induced breakdown spectroscopy". In: *Journal of the Brazilian Chemical Society* 18 (2007), pp. 463–512.

[5] Sergio Musazzi and Umberto Perini. "Laser-induced breakdown spectroscopy". In: *Springer Series in Optical Sciences* 182 (2014).

[6] David A Cremers and Leon J Radziemski. *Handbook of laser-induced breakdown spectroscopy*. John Wiley & Sons, 2013.

[7] RPMC Lasers, Inc. *OEM Fiber Lasers For Industrial Laser Induced Breakdown Spectroscopy*. July 5, 2023. URL: https://www.rpmclasers.com/blog/oem-fiber-lasers-industrial-laser-induced-breakdown-spectroscopy/.

[8] Muskan Shaik. *Physics and radio electronics - Laser*. July 5, 2023. URL: https://www.physics-and-radio-electronics.com/physics/laser/methodsofachievingpopulationinversion.html.

[9] Sébastien Forget François Balembois. *Optics 4 engineers - Q-switching*. July 5, 2023. URL: http://www.optique-ingenieur.org/en/courses/OPI_ang_M01_C01/co/Contenu_12.html.

[10] Andor. *Overview of Echelle Spectrograph - Flexible Spectroscopy Tool*. July 6, 2023. URL: https://andor.oxinst.com/learning/view/article/echelle-spectrographs-a-flexible-tool-for-spectroscopy.

[11] Jagdish P Singh and Surya N Thakur. *Laser-induced breakdown spectroscopy*. Elsevier, 2020.

[12] Siying Chen et al. "Approximate Voigt function formula for laser-induced breakdown spectroscopy fitting". In: *Applied Optics* 60.14 (2021), pp. 4120–4126.

[13] Christian Hill. *Learning Scientific Programming with Python*. July 10, 2023. URL: https://scipython.com/book/chapter-8-scipy/examples/the-voigt-profile/.

[14] Elisabetta Tognoni and Gabriele Cristoforetti. "Signal and noise in laser induced breakdown spectroscopy: an introductory review". In: *Optics & Laser Technology* 79 (2016), pp. 164–172.

[15] Johann Lohninger. *Fundamentals of Statistics*. July 10, 2023. URL: http://www.statistics4u.com/fundstat_eng/index.html.

[16] Sangwoo Yoon, Hae-Woon Choi, and Joohan Kim. "Analysis of changes in spectral signal according to gas flow rate in laser-induced breakdown spectroscopy". In: *Applied Sciences* 11.19 (2021), p. 9046.

[17] Peter Filzmoser. *Advanced Methods for Regression and Classification - Lecture Notes*. Oct. 1, 2022.

[18] Chirag Goyal. *Feature Transformations in Data Science: A Detailed Walkthrough*. Mar. 6, 2021. URL: https://www.analyticsvidhya.com/blog/2021/05/feature-transformations-in-data-science-a-detailed-walkthrough/.

[19] Songhao Wu. *Is Normal Distribution Necessary in Regression? How to track and fix it?* May 15, 2020. URL: https://towardsdatascience.com/is-normal-distribution-necessary-in-regression-how-to-track-and-fix-it-494105bc50dd.

[20] Roberto Reif. *Importance of Feature Scaling in Data Modeling*. Dec. 16, 2017. URL: https://www.robertoreif.com/blog/2017/12/16/importance-of-feature-scaling-in-data-modeling-part-1-h8nla.

[21] In-Kwon Yeo and Richard A Johnson. "A new family of power transformations to improve normality or symmetry". In: *Biometrika* 87.4 (2000), pp. 954–959.

[22] Abhay Parashar. *Feature Transformation and Scaling Techniques*. Nov. 28, 2020. URL: https://pub.towardsai.net/feature-transformation-and-scaling-techniques-f9645cb538e.

[23] Amiya Ranjan Rout. *GeeksforGeeks - Spearmans Rank Correlation*. July 10, 2023. URL: https://www.geeksforgeeks.org/spearmans-rank-correlation/.

[24] Iain Pardoe; Laura Simon; Derek Young. *Stat 501 - Course Notes. Lesson 2: SLR Model Evaluation*. PennState Eberly College of Science. July 11, 2023. URL: https://online.stat.psu.edu/stat501/lesson/2.

[25] Zach Bobbit. *Statology - Standard Error of Regression Slope*. Sept. 30, 2021. URL: https://www.statology.org/standard-error-of-regression-slope/.

[26] David Prochazka et al. "Machine learning in laser-induced breakdown spectroscopy as a novel approach towards experimental parameter optimization". In: *Journal of Analytical Atomic Spectrometry* 37.3 (2022), pp. 603–612.

[27] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[28] The pandas development team. *pandas-dev/pandas: Pandas*. Version 2.0.1. Apr. 2023. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[29] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[30] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[31] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[32] Matthew Newville et al. "LMFIT: Non-linear least-square minimization and curve-fitting for Python". In: *Astrophysics Source Code Library* (2016), ascl–1606.

[33] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[34] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[35] Prof. Shalabh. *Regression Analysis. 12. Polynomial Regression Models*. Ed. by Indian Institute of Technology Kanpur. Aug. 3, 2023. URL: https://home.iitk.ac.in/~shalab/regression/Chapter12-Regression-PolynomialRegression.pdf.