Machine Learning Methods for the Raw Data Analysis of Cryogenic Dark Matter Experiments

Felix Wagner, B.Sc.

Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften

Atominstitut, Technische Universität Wien

A thesis submitted for the degree of

Master of Science

under the guidance of

Dr. Florian Reindl
Priv.-Doz. Dr. Wolfgang Waltenberger
Univ.-Prof. Dr. Jochen Schieck

2020

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

I finished my thesis during the Covid-19 outbreak in 2020 and dedicate it to those, who kept our infrastructure running throughout this crisis.
Thank you for risking your health for us.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Abstract

The CRESST experiment uses cryogenic particle detectors to measure low-energy particle interactions, which allows to probe sub-GeV/c^2 dark matter. The high sensitivity of the experiment requires a sensible analysis of the raw data, which consists of one-dimensional voltage pulse traces. Particle recoils create events with a characteristic pulse shape, by which they can be identified between other, non-physical events. In this work, we present novel analysis tools based on machine learning. We introduce a method for the simulation of labeled data sets and train an LSTM neural network for the pulse height reconstruction of low energetic events. Our model beats the state-of-the-art method and precisely determines pulse heights between 0.6 mV and 6.8 mV, which corresponds to signal-to-noise ratios between 0.2 and 1.5. Furthermore, we are able to cluster the raw data distribution in event classes with similar properties by the use of unsupervised methods, including PCA, t-SNE and autoencoders. In addition, we train an LSTM model for the classification of simulated events and reach a discrimination threshold of 3.5 mV, which is an order of magnitude lower than that of the previously used method. Our model performs well on a small, hand-labeled set of raw data events. These methods are not only applicable to CRESST-data, but to all similar cryogenic experiments.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Kurzfassung

Das CRESST-Experiment verwendet kryogene Teilchendetektoren zur Messung niederenergetischer Wechselwirkungen, mit denen sub-GeV/c^2 dunkle Materie Kandidaten untersucht werden können. Die hohe Empfindlichkeit des Experiments erfordert eine sensible Analyse der Rohdaten, die aus Zeitreihen von Spannungswerten bestehen. Teilchenrückstöße erzeugen Ereignisse mit einer charakteristischen Pulsform, anhand derer sie von anderen, nicht physikalischen Ereignissen unterschieden werden können. In dieser Arbeit stellen wir neuartige Analysewerkzeuge vor, die auf maschinellem Lernen basieren. Wir stellen eine Methode zur Simulation von gelabelten Datensätze vor und trainieren ein neuronales LSTM-Netzwerk für die Rekonstruktion der Pulshöhe von niederenergetischen Ereignissen. Unser Modell übertrifft die State-of-the-Art Methode und bestimmt präzise Pulshöhen zwischen $0,6$mV und $6,8$mV, was einem Signal-Rausch-Verhältnis zwischen $0,2$ und $1,5$ entspricht. Mit PCA, t-SNE und Autoencodern erreichen wir eine Clusterbildung der Rohdatenverteilung in Ereignisklassen mit ähnlichen Eigenschaften. Wir trainieren ein LSTM-Modell für die Klassifizierung simulierter Ereignisse und erreichen eine Unterscheidungsschwelle von $3,5$mV, die um eine Größenordnung niedriger ist als die der zuvor verwendeten Methode. Unser Modell liefert gute Ergebnisse an einem kleinen, handbeschrifteten Rohdatensatz. Unsere Methoden sind nicht nur für CRESST-Daten anwendbar, sondern für alle ähnlichen kryogenen Experimente.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Contents

Introduction **1**

1 A Brief Introduction to Dark Matter **4**

1.1 Experimental Evidence for the Existence of Dark Matter 5

1.2 Particle Candidates 9

1.3 Experimental Search for Dark Matter 11

2 The CRESST Experiment **14**

2.1 The Setup of the Experiment 16

2.2 Cryogenic Particle Detector Modules 17

2.3 Analysis of CRESST Data 20

2.3.1 Raw Data Analysis 21

2.3.2 High Level Analysis 25

3 Data Simulation **27**

3.1 The Pulse Shape Model 29

3.2 Simulated Noise Baselines 32

3.3 The Simulated Recoil Data Set 36

3.4 High Energy Recoils 38

4 Pulse Height and Onset Timing Regression **45**

4.1 The Optimum Filter 46

4.2 The Standard Event Fit 51

4.3 Deep Regression 56

4.3.1 Pulse Height Regression 59

4.3.2 Onset Timing Regression 62

4.4 Results on Real Data 65

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

5 Unsupervised Techniques: Dimensionality Reduction, Anomaly Detection **68**
5.1 The Carrier Event 69
5.2 Dimensionality Reduction 71
5.2.1 Principal Component Analysis 71
5.2.2 t-distributed Stochastic Neighbor Embedding 74
5.3 Anomaly Detection 77
5.3.1 The Standard Event Fit 78
5.3.2 Auto-Encoding Neural Networks 80

6 Supervised Event Classification **85**
6.1 Classification of Carrier Recoils 86
6.2 Classification of Noise Baselines 89
6.3 Application to Measured Data 92

Conclusion **96**

A An Introduction to Deep Learning **99**
A.1 Feed Forward Networks 100
A.2 Convolutional Neural Networks 102
A.3 The ResNet 103
A.4 Long Short-Term Memory Networks 105
A.5 Autoencoders 107
A.6 Generative Adversarial Networks 108

B Simulated Noise Baselines with Generative Adversarial Networks **111**

Bibliography **115**

Index **122**

Acknowledgements **124**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Introduction

Today we observe a discrepancy between the visible matter density in our universe and the matter density that is needed to explain various phenomena on cosmological and galactic scale [1–3]. The most plausible explanation for this discrepancy is the existence of a non-visible form of matter, the so-called dark matter, that makes up about 85% of the total matter density. The actual nature of dark matter, especially if it features a particle-character like ordinary matter, is one of the biggest open questions in modern physics.

The CRESST (Cryogenic Rare Event Search with Superconducting Thermometers) experiment attempts to detect dark matter particles directly from the galactic dark matter halo, through which our planet constantly moves. The unique characteristic of this experiment is the use of scintillating crystals as cryogenic detectors, that enable the detection of Sub-GeV/c^2 particles and the efficient discrimination between different types of particle recoils [4]. This concept proved to be efficient, so that today two spin-off experiments, NUCLEUS and COSINUS, plan on using the same detection technology [5, 6].

As CRESST is one of the most sensitive experiments, a sensible data analysis is crucial. The detector data consists of voltage pulse traces, that bear the characteristic shape of a pulse for recoil events within the detector crystal [7]. These pulse-shaped events have to be triggered from the continuous data stream, the corresponding recoil energy has to be determined and non-physical events, as detector artifacts, have to be discarded. The output of the raw data analysis is the time and energy information of all recoil events, from all detectors of the experiment. From this information, dark matter exclusion limits are calculated. The quality of each step of the raw data analysis directly affects the quality of the exclusion limit and therefore the quality of the results of the whole experiment. The raw data analysis is partially automatized by the use of matched filters, fitting techniques and especially designed, parametric cuts. Combined with the detector concept, this makes trigger thresholds down to 30.1 eV

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

possible, with a survival rate of events originating from recoils within the absorber crystal between 50% and 65%, strongly depending on the energy of the event [8].

In the past decade, machine learning algorithms experienced a phase of rapid development. These methods are unique compared to other statistical methods, as they learn from provided data without explicit instructions. They often find efficient strategies to solve classification and regression problems for high dimensional and highly correlated data. Especially deep learning methods, i.e. machine learning with neural networks, produce astonishing results in the realm of image- and audio recognition and generation [9]. Modern machine learning methods have already, and will even more in the coming years, open new paths for data-driven research areas in the natural sciences [10]. There are already applications for the search for dark matter and comparable rare-event searches, from experimental, as well as from theoretical side [11–13].

Deep learning methods were already investigated for a discrimination task within the CRESST data analysis, namely to identify carrier events, a specific type of artifacts [14]. In our work, we further study the application of neural networks for the discrimination of different event classes, we provide several methods for the visualization of the raw data distribution and present a method to precisely reconstruct the energy of events. By this, we point out two different approaches to effectively lower the trigger threshold and one to raise the exposure time due to an improved event discrimination. The structuring of this thesis is explained in the following.

In the first chapter we introduce the experimental evidence for the existence of dark matter. Furthermore, we discuss the requirements for suitable dark matter particle candidates and provide an overview of different types of dark matter experiments.

Thereafter, in chapter two, we examine the setup and detector modules of the CRESST experiment, especially the TUM40 module, the best performing and well-studied CRESST-II module. We introduce the working principle of transition edge sensors (TES) and give a short introduction into CRESST data analysis.

In chapter three we introduce a method to simulate recoil events and noise baselines and discuss the challenges of the energy reconstruction for saturated, high energetic pulses.

We study the optimum filter, the standard event fit and a deep learning regression model as methods for the reconstruction of pulse heights in chapter four. We compare the results of these methods and test them on exemplary chosen raw data events.

In chapter five we study four unsupervised techniques, i.e. methods that do not rely on a labeled data set. The principal component analysis (PCA) and the t-distributed

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

stochastic neighbor embedding reduce the dimensionality of a given data distribution to make an exploratory data analysis and visualization possible. The standard event fit and the autoencoder could serve as quality cuts by discarding events whose voltage trace shape diverges from the characteristic pulse shape.

In chapter six we study supervised classification of simulated absorber events, carrier events and noise baselines with neural networks. We provide an evaluation on a small, hand-labeled sample of raw data events and discuss possible implications for the dark matter exclusion limit.

For readers not familiar with deep learning, we recommend the reading of the condensed introduction to neural networks that we provide in appendix A, parallel to, or before, the examination of chapter 4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Chapter 1

A Brief Introduction to Dark Matter

The total energy density in our universe is mainly made up of four parts: Visible, also called *baryonic* matter, *dark matter*, radiation and *dark energy*. Within the Λ-CDM (Lambda cold dark matter) model, the current standard model of Big Bang cosmology, the contribution of each of the constituents scales differently with the expansion of the universe and we are able to give estimates on the contribution of each part [15, 16]. Today, the relevant parts of the energy density are visible and dark matter and dark energy. Radiation, mainly through the cosmic microwave background, makes up only a very small part ($\sim 0.01\%$) of the total energy density, however, this was different in the early universe, when radiation made up most of it's energy density. Dark energy is a still not understood constant energy of the empty space, described by the cosmological constant in Einsteins equations, that is responsible for the current acceleration in the expansion of our universe [17]. Today it makes up about 70% of the energy density. The contribution from matter splits up into visible, baryonic matter and dark matter. Baryonic matter – which is mainly made up of large clouds of helium and hydrogen, but also includes stars and galaxies – accounts for roughly 5% and is relatively well understood through the standard model of particle physics. This is not the case for dark matter which makes up a by-far bigger part of about 25% of the universes energy density. Other than dark energy, dark matter interacts with baryonic matter through gravitation and possibly the weak force [18]. From this interaction it is even possible to calculate a dark matter density on a galactic scale. It is, however, still an unanswered question if dark matter also consists of particles, as no dark matter particle has been measured yet. This chapter starts with the discussion of experimental evidence for the existence of dark matter on different scales. In the second section, the WIMP (Weakly Interacting Massive Particle) is introduced, a

simple and suitable model for a dark matter particle candidate. The chapter concludes by giving an overview over the various ways to search for dark matter experimentally and a short discussion of the current status of this experimental search.

1.1 Experimental Evidence for the Existence of Dark Matter

The first to notice a form of non-visible matter was the Swiss astronomer Fritz Zwicky in the year 1933, during his observations of the velocity distribution of galaxies within the coma cluster at the Californian Institute of Technology. He used the Virial theorem to estimate a mean mass density that was up to 400 times larger than the visible mass distribution within the studied galaxies [19]. Today we know of a wide range of further evidence on cosmological and galactic scale, and even within individual galaxies.

To start with the cosmological evidence, a precise measurement of the dark matter density today is possible via the *cosmic microwave background (CMB)*. The CMB was discovered in 1964 by Arno Penzias and Robert Wilson, when they measured a strange radiation, that seemingly arrived on earth isotropically from all sides [20]. At first they thought the signal came from a bird family that had nested in there antenna - however, the signal remained after they removed the birds and all their residuals, which they kindly called "white dielectric material" [21].

Today we know that the cosmic microwave background radiation was emitted at the stage in the evolution (and the corresponding temperature decrease) of the universe, when the mean energy density dropped below 13.6 eV - which is the ionization energy of hydrogen. This happened about four hundred thousand years after the Big Bang and by this, the universe became transparent. The universe was still very young back then, considering its current age of about 13.8 billion years. It had only about 1/1000 of its current size.

The CMB shows the best black-body spectrum ever observed in nature. It is depicted in Fig. 1.1. As it cooled down proportionally to the expansion of the universe, its current intensity-spectrum corresponds to a temperature of 2.726 K. Down to a scale of 10^{-5} it is isotropic, i.e. no matter in which direction your antenna points, it stays the same. However, on smaller scales it's temperature variations are the best tool for precision measurements in cosmology (see Fig. 1.1, red - higher values, blue - lower values), as they provide information about matter density fluctuations in the early universe. The clumping of matter in the primordial plasma leads to a counter-pressure of radiation within the regions with higher density. Opposed to the gravitational pull

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

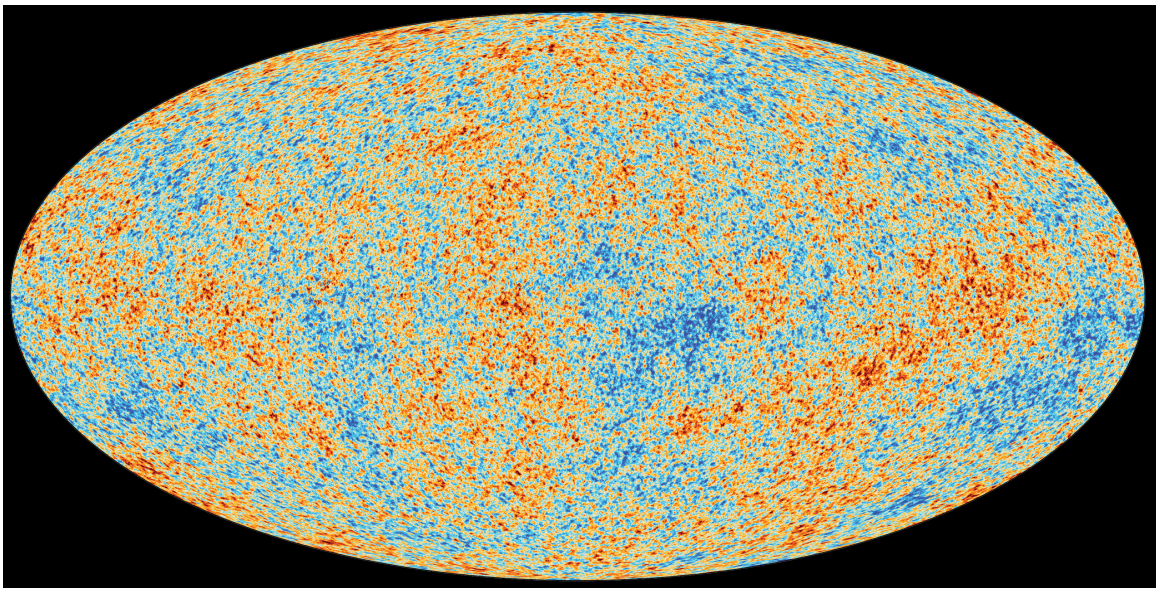
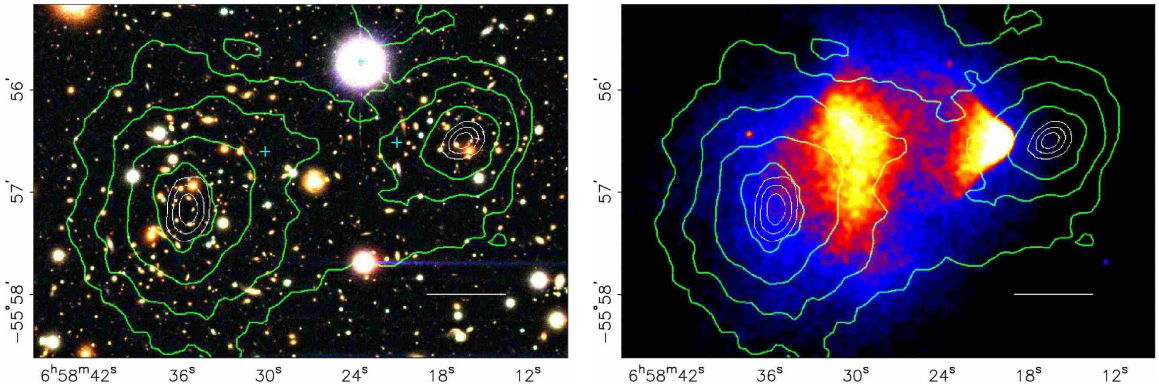


Figure 1.1: The anisotropies of the cosmic microwave background, or CMB, as observed by ESA's Planck mission. The CMB is a snapshot of the oldest light in our cosmos, imprinted on the sky when the Universe was just 380000 years old. It shows tiny temperature fluctuations (red - higher values, blue - lower values) that correspond to regions of slightly different densities, representing the seeds of all future structure: the stars and galaxies of today. This image is based on data from the Planck Legacy release, the mission's final data release, published in July 2018 [22].

that comes from the higher density, this causes oscillations in the plasma, which can be seen in the angular power spectrum of the CMB today. Dark matter, which does not interact with electromagnetic radiation, only leads to a higher gravitational pull, and therefore produces a characteristic signature. From this, the total matter density in the universe was estimated to 26.9% of the total energy density, while the baryonic matter density was estimated to only 4.6% of the total energy density.

In addition, it turns out that the *structure formation* of the universe itself was only possible due to the existence of non-baryonic, non-visible matter. Under the assumption of a Big Bang, that brought the universe into existence, we first have to clarify, why the so produced plasma even developed inhomogeneities, that later became stars and galaxies, in the first place. In fact, this happened because of quantum fluctuations, that occurred sometime between the first 10^{-38} s and 10^{-14} s after the big bang. This epoch is called *inflation* and it's characteristic is an exponentially fast expansion. A quantum fluctuation occurring in this epoch was amplified through the expansion, such that it got large enough to cause a lasting inhomogeneity in the plasma. Due to the gravitational pull of the dense region, those clumps grew. However,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 1.2: Shown above in the left panel (a) is a color image from the Magellan images of the merging cluster 1E0657-558, with the white bar indicating 200 kpc (1 pc $\approx 3 \cdot 10^{16}$ m) at the distance of the cluster. In the right panel (b) is an image of the cluster from the Chandra X-ray observatory. Shown in green contours in both panels is the reconstruction of the mass distribution by gravitational lensing. We observe, that the centers of matter that emits radiation does not match with the centers of the mass distribution [24].

this led, as mentioned above, to a counter pressure of radiation. In the early universe, the counter pressure was strong enough to disperse the clumps again - therefore it was not possible for baryonic matter to build clumps. This was different for dark matter, which does not feel the counter pressure of the radiation. Without dark matter, our universe would therefore have started to clump at a much later time as it actually did. This fascinating realization was supported by several, large computer simulations, e.g. the so-called Millennium Simulation [23].

Besides the CMB and the structure formation, cosmology provides a third evidence, that lies even further in the past - namely in the time of *Big Bang nucleosynthesis*, about three minutes after the Big Bang. This was when protons and neutrons dropped out of equilibrium and the first atomic nuclei formed. These first nuclei were only very light elements, up to ^{7}Li. From the number of light elements that we observe today, a strong constraint can be put on the number of baryons at the time of nucleosynthesis. This constraint matches very well with the prediction of a current baryonic matter energy density of about 4% of the total energy density.

On a smaller scale, namely the scale of galaxy clusters, there are three phenomena that hint towards the existence of dark matter. One, the observation of velocity distributions, was introduced at the very beginning of this chapter. The second is provided due to the *gravitational lensing* effect. A gravitational lens is a heavy object, often a galaxy cluster, that bends light in its gravitational field, leading to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

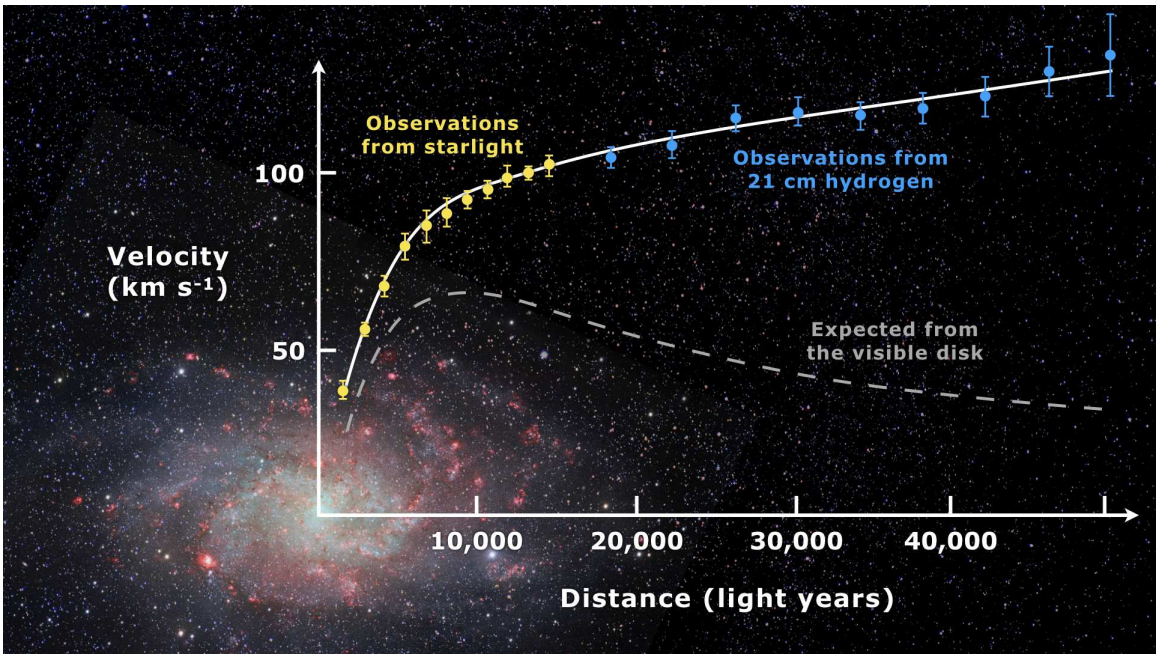


Figure 1.3: Rotation curve of spiral galaxy Messier 33 (yellow and blue points with error bars), and a predicted one from distribution of the visible matter (gray line). The discrepancy between the two curves can be accounted for by adding a dark matter halo surrounding the galaxy [25, 26].

distortions and even double-vision of objects. From observations of the strength of the bending effect, the mass of the gravitational lens can be calculated. Also here, a strong disagreement between the calculated and the visible mass exists. A third, and direct, observation of dark matter was done on the so-called *bullet cluster*, which is depicted as an optical image in Fig. 1.2a. It actually consists of two galaxy clusters, that collided around 150 million years ago. The green contour lines mark the gravitational potential calculated from gravitational lensing. Two distinct centers of mass are visible, corresponding to the two galaxy clusters. In Fig. 1.2b, we see an X-ray image of the bullet cluster, the hot (red) regions correspond to the location of hot gas, which usually makes up the most of the baryonic matter in galaxy clusters. We can see a clear discrepancy between the location of the hot gas, the stars and the centers of mass. This also requires a form of weakly-interacting, non-visible, cold matter to explain the centers of mass of the two galaxy clusters. The displacement of barionic and dark matter was caused by the different scattering strengths in the collision of the two clusters.

As a last one, we also want to introduce the evidence provided by the *rotation curves* of single galaxies. In Fig 1.3 the M33 galaxy is shown. The dashed line shows the velocity distribution, that is expected from the visible matter-distribution, and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

falls off proportionally to $\sim 1/\sqrt{r}$. The continuous lines shows the actually observed velocity distribution, which stays roughly flat. The discrepancy can be explained by a spherical *dark matter halo*, that surrounds the visible galaxy and is several times larger than it. This flat form of velocity distribution occurs for a large number of galaxies, the milky way being one of them.

Having introduced the various hints toward the existence of dark matter, we continue in the following part with a summary of all requirements a hypothetical dark matter particle would have to fulfill and introduce some candidates for such particles.

1.2 Particle Candidates

Before we come to the discussion of general requirements for non-baryonic matter candidates, we want to discuss if any of our standard model particles could explain the observations. There are indeed three particles that interact only weakly with baryonic matter and decouple from the plasma about the same time as dark matter does: The *neutrinos*. However, at the point of decoupling neutrinos move still with relativistic speed. Relativistic (*hot*) dark matter would favor the formation of large clusters, the size of superclusters. Non-relativistic (*cold*) dark matter favors the formation of smaller clusters, like galaxies. The formation of superclusters seems to be a rather new development, other than the formation of galaxies. Therefore neutrinos can be ruled out as dark matter candidates, because they are simply not heavy enough.

Explainable within the standard model, but non-visible, there are also *MACHOs (massive compact halo object)* and *black holes*, that do contribute gravitational effects in the universe [27]. Their contribution is however too small to explain the observed phenomena. We can therefore rule out any already known baryonic matter particles to explain the dark-matter phenomena.

From the experimental evidence that was introduced in the former section, we know that a dark-matter particle has to fulfill the following requirements: It has to be cold, i.e. already non-relativistic at the time of decoupling, maximally weakly interacting with baryonic matter and stable on time scales of the age of the universe. The force through which dark matter interacts with baryonic matter does not necessarily have to be the weak force from the standard model of particle physics, however, it may not be significantly stronger than it.

Two matching candidates e.g. are *massive sterile neutrinos*, right-handed neutrinos that do not interact weakly, and axions, hypothetical particles corresponding to an

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

additional symmetry that would solve the strong CP problem [28]. Particles likes these, that only interact through gravity, not through the weak force, would very likely never get detected by any experiment. Luckily, there are some indications that lead to the suspicion, that dark matter could very well interact with baryonic matter through the weak force.

The *WIMP (weakly interactive massive particle)*, named after it's assumed properties, is a stable, or very long-lived, particle that fulfills exactly the above described requirements, i.e. interacting only with the weak force, being relatively heavy and being per definition non-baryonic. The two most important properties of a particle like this would be its mass and its interaction cross section. The cross section can be estimated from cosmological arguments, namely from the annihilation rate of particles and anti-particles that lead to the current density - as well as from Fermi's Golden Rule, by assuming a particle mass of $\sim \mathcal{O}(10\ \mathrm{GeV/c^2})$, typical for the weak interaction. The estimation for the cross section is around $\sim \mathcal{O}(10^5\ \mathrm{pb\ km/s})$. The fact, that a particle with this mass and cross section would drop out of thermal equilibrium with the primordial plasma at the right time and then decay with a proper rate to reach the current dark matter density - and therefore explain all experimental observations, is not at all trivial and reverently called the WIMP miracle.

There is a valid explanation for the existence of a WIMP trough a famous extension of the standard model: *Supersymmetry (SUYS)*. This theory assumes an underlying symmetry between fermions and bosons in the standard model, that would assign a partner-boson to every fermion and vice versa. Those particles have to conserve a quantity, the so-called R-parity, that prohibits the *lightest supersymmetric particle (LSP)* from decaying into standard model particles. It would therefore be stable and a good fit for a dark-matter particle [29]. There are many more theories, that predict suitable candidates. However, neither supersymmetry, nor any of the others are experimentally verified today.

Aside from theories that predict dark matter particles, there are theories that try to explain the rotation curves of galaxies with *modifications of Newtonian dynamics (MOND)* [30]. These however cannot explain all phenomena, e.g. those from cosmology and gravitational lensing.

We will in this work stick with the assumption of a WIMP, as we described it above, as the dark matter particle. This is also the most common assumption for the experimental search for dark matter, which we will discuss in the following chapter.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

1.3 Experimental Search for Dark Matter

Many experiments try today to verify the particle character of dark matter, i.e. to measure a dark matter particle. These experiments can be classified by their detection strategy, i.e. direct or indirect detection and production at colliders.

This work focuses on the *direct detection* of dark matter, we will only discuss the other two approaches briefly. First, there are experiments that attempt to produce dark matter particles from the annihilation of standard model particles at particle colliders, e.g. at the LHC. These should then leave characteristic traces, namely missing energy in the detectors. This however has several disadvantages, most importantly the lack of information about the life time of dark matter particles, but the lack of information about the coupling of dark matter to leptons at hadron colliders. It does however provide a good sensitivity for low-mass dark matter particles. Differently to that, indirect detection experiments provide a high sensitivity towards high-mass dark matter particles. They aim to measure the decay products of interactions of dark matter particles and standard model particles, e.g. by the use of satellites (Fermi-LAT, ...) or Cherenkov teleskopes (MAGIC, HESS, CTA, ...). These aim to detect dark matter signatures e.g. in the gamma flux from dwarf spheroidal galaxies or the neutrino flux from the galactic center.

Direct detection experiments aim to detect interactions of dark matter particles with standard model particles directly in the detector material. As earth itself is located in the galactic dark matter halo of the milky way, the scattering of dark matter particles on earth-based detector targets should happen frequently, depending on several factors: First, the number density of the target nuclei in the detector. Second, the velocity distribution and density of dark matter particles at the detectors location. This distribution can be estimated from the milky-way's velocity distribution. The third factor is the interaction cross section and the mass of dark matter particles. The cross section also depends on the material of the detector itself. Possible further dark matter particle properties, e.g. the spin, may play a role, but are neglected under standard assumptions. A low expected scattering rate and recoil energies in the order of magnitude of $\mathcal{O}$(keV) require highly sensitive detectors, high exposure, typically measured in kg days, and low background. There are different approaches to fulfill these requirements. The CRESST experiment, which is discussed in chapter 2 in detail, aims for a low recoil energy threshold by the use of cryogenic, light-weight detectors. This especially enables the measurement of light dark matter candidate particles [31].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

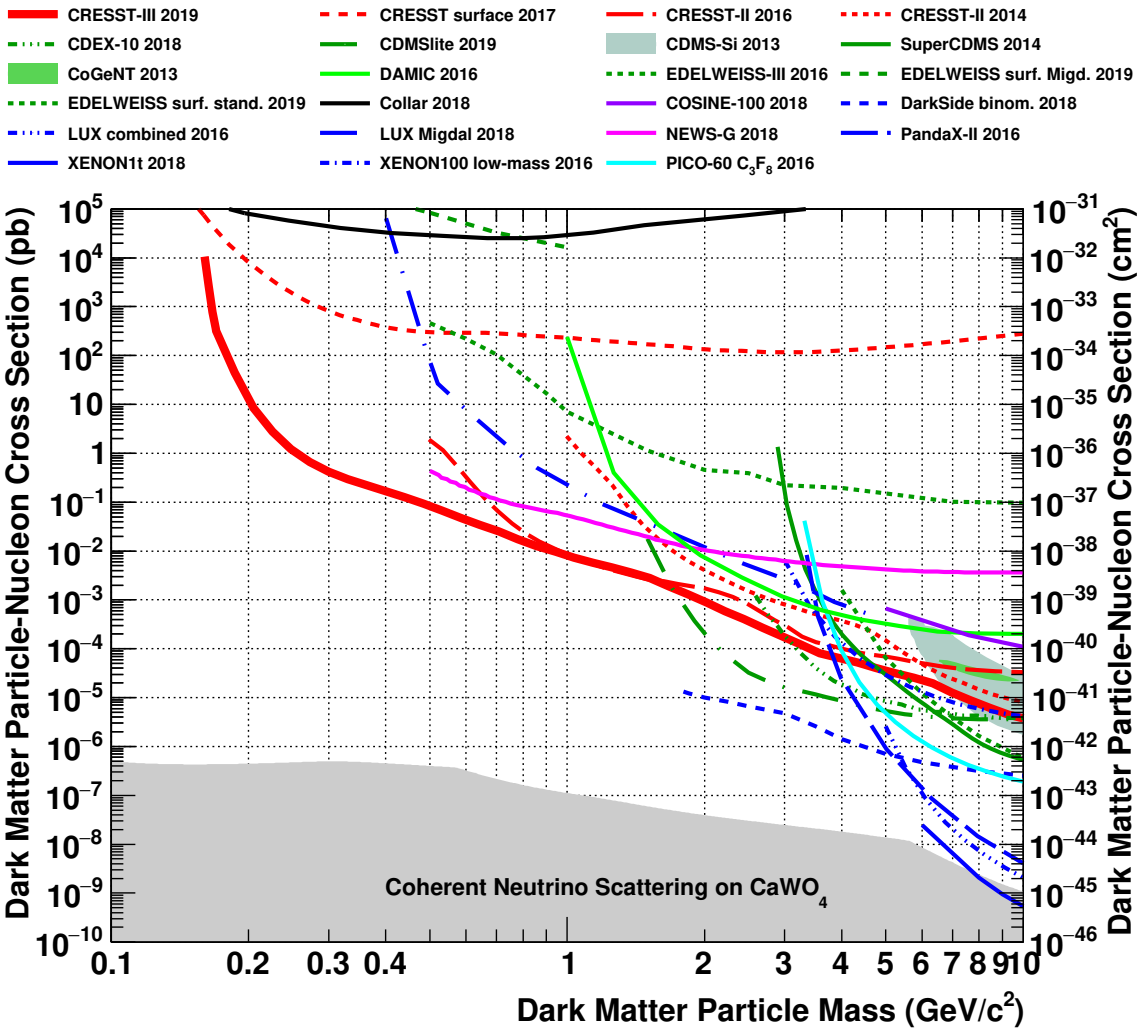


Figure 1.4: Dark matter exclusion limits from the CRESST experiment (red), liquid noble gas experiments (blue) and other cryogenic experiments (green). The area above the line is excluded with 90% confidence to not contain any dark matter particles, with data from the corresponding experiment. The best limit for low mass dark matter is achieved by the CRESST-III experiment (red, continuous). The gray region corresponds to the parameter space that is, for $CaWO_4$-based experiments, concealed by the coherent neutrino-nucleus scattering [8].

The typical way to quantify and compare results from dark matter experiments is the calculation of *exclusion limits*. These are visualized in a two dimensional plot, one dimension being the dark matter mass, the other being the dark matter cross section. Lines in this plot correspond to limits, while positive signals are shown as allowed regions. In Fig. 1.4, the exclusion limits, calculated for data from different experiments, are shown. The most recent results from the CRESST-III experiment (red, continuous) impose currently the best limit on low mass dark matter, under

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

standard assumptions. There is however a more recent analysis of XENON1t data, that involves the Migdal effect, and therefore calculates an even lower limit for low masses [32]. The best limit that could be achieved with data from the CRESST-II experiment was recorded with the Lise detector (red, widely dashed). The data, that is used in this thesis, is also from the CRESST-II experiment, but was recorded with the TUM40 detector (see chapter 2).

This was only a very brief introduction to dark matter. There is much more to say and know that would exceed the scope of this thesis. Let us therefore refer the interested reader to Ref. [33, 34] for deeper insights. The CRESST experiment is one direct detection experiment, on which this thesis is focused. It is discussed in the following chapter in detail.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Chapter 2

The CRESST Experiment

In the previous chapter, we introduced dark matter and one promising particle candidate: The WIMP. CRESST (Cryogenic Rare Event Search with Superconducting Thermometers) is an experiment with the goal to directly detect WIMPs via their scattering off target nuclei in cryogenic, scintillating detectors. It is located in the *Laboratori Nazionali del Gran Sasso (LNGS)* in the Abruzzi mountains in central Italy. The experimental collaboration has contributors from seven research institutes, namely the Max-Planck-Institute for Physics Munich, the University of Oxford, the Technical University Munich, the University Tübingen, the Istituto Nazionale di Fisica Nucleare (INFN) Italy, the Institute of High Energy Physics of the Austrian Academy of Sciences and the Technical University Vienna. The construction on the experiment started in 1995, first data was taken in 2000. In its early stage, the detectors were not yet made of scintillating material, therefore it was difficult to discriminate between nuclear recoils and electromagnetic events, because their respective signatures on thermometers alone are hardly distinguishable [4]. In 2009, the experiment was upgraded to CRESST-II phase 1, from which point on mainly scintillating $CaWO_4$ crystals were used. It is possible to classify different event types with the ratio between produced scintillation light and temperature increase [35]. In 2013 the detectors were upgraded and the following three years are called phase 2 [36]. In 2016 the experiment was again upgraded and is today called CRESST-III. It is now optimized for the detection of light dark matter particles. Indeed, the most current results of CRESST-III impose, under standard assumptions, the strongest limit on Sub-GeV/c^2 dark matter of all direct detection experiments [8].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

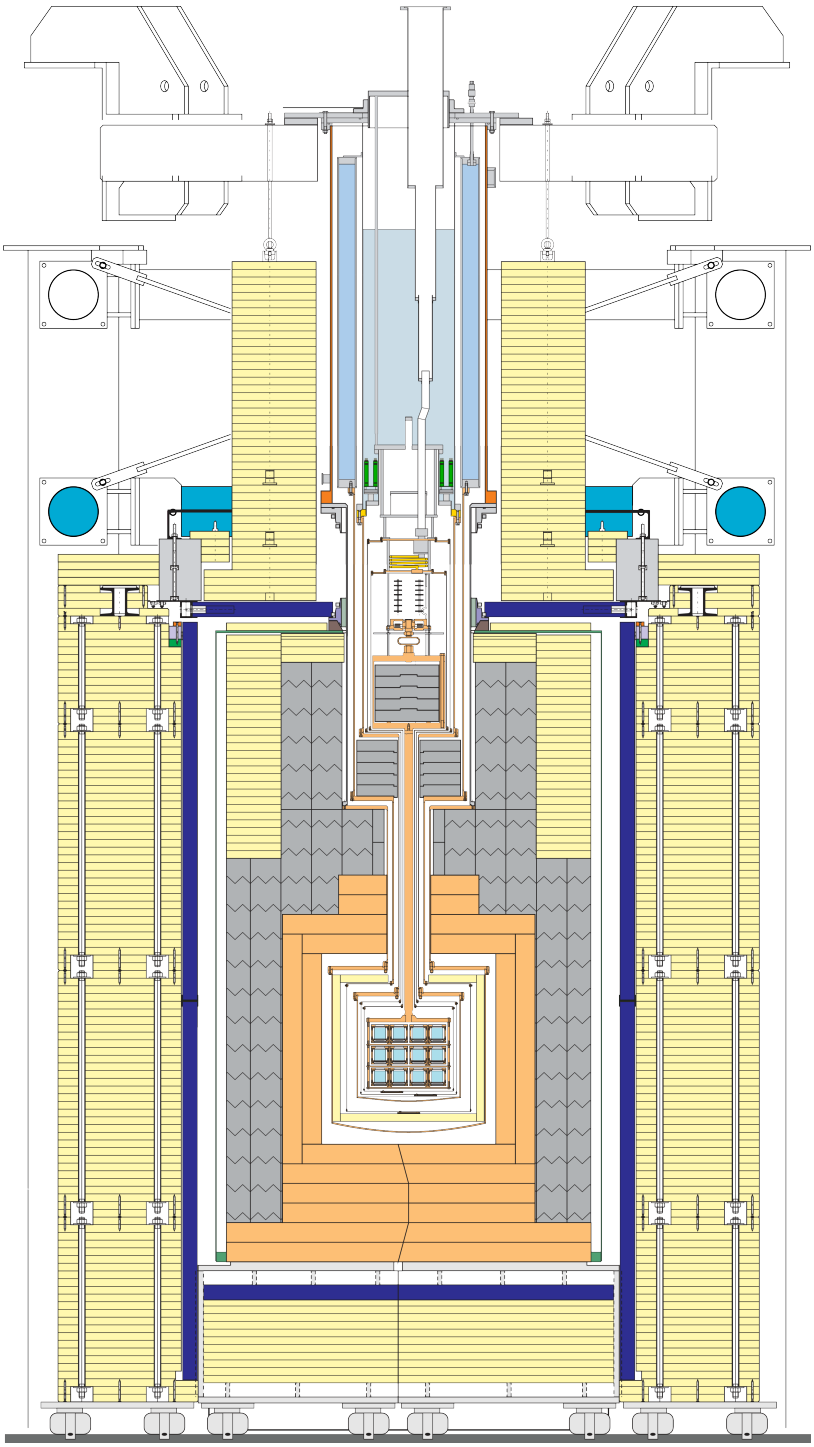


Figure 2.1: The setup of the CRESST experiment. The largest part is the shielding, consisting of layers of lead (dark grey), copper (bronze), polyethylene (yellow) and scintillating plastic panels (dark blue). In the center of the shielding, the carousel with the detector modules is mounted. A copper rod, called cold finger, connects the carousel with the cryostat, which is located above the carousel and outside the shielding. The purpose of the different layers is explained in section 2.1. Picture adapted from Ref. [37].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

This thesis works with data from phase 2 of the CRESST-II experiment, but also introduces some of the novelties that came with CRESST-III. In this chapter we cover the general setup of the experiment, followed by the detector modules. We end with an introduction of the current data analysis tools. Possible improvements for these tools are the main focus of this thesis.

2.1 The Setup of the Experiment

The results of Rare Event Searches are high precision measurements and sensitive to background radiation. Such experiments are therefore conducted underground. The LNGS, the location of the CRESST experiment, is located under 1400 m of rock which equals 3800 m water, absorbs most high energy cosmic radiation and reduces muon flux by a factor of 10^6 [31]. However, there are still environmental influences from which we have to shield the detectors. These are mainly due to remaining muons and the natural radioactivity of the surrounding rock and the lead within the shielding.

The setup of the experiment is depicted in Fig. 2.1. At the very center of the roughly 6 m high and 3 m wide construct is the so-called *carousel*, in which the detectors are mounted. The detectors need to be operated at cryogenic temperatures, more precisely at temperatures below 10 mK. These temperatures are accomplished with a commercial dilution refrigerator circulating a He^3/He^4 mixture. The cryostat and the tanks of liquid helium and nitrogen are mounted about two meters above the carousel and connected to it by a copper rod, called the *cold finger*.

The largest part of the setup is the *shielding*, which protects the experiment from several influences:

- The dark blue parts in Fig. 2.1 are plastic scintillator panels, that act as an active *muon veto*. They tag muons that were not absorbed in the rock, such that they can get filtered automatically in the data analysis.

- The natural radioactivity in the rock originates mainly from ^{4}K, ^{238}U, ^{232}Th and their respective decay chains. This requires a shielding against the gamma background, which is realized by a 20 cm thick layer of 24 tons of lead (dark gray areas in Fig. 2.1).

- However effective lead might be against the gamma background, the material itself is not radio-pure. Therefore a layer of copper is placed inside the lead layer (bronze in Fig. 2.1). Copper can be produced radio-purely, which is why it is also used for the carousel and the cold finger.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

- *Neutrons* are another background source, as they produce the same signature on the detector as a WIMP would. Polyethylene is an efficient moderator for neutrons because of its high number of light hydrogen atoms. Two layers of polyethylene are placed outside the lead shielding and inside the copper shielding, respectively. They provide protection from neutrons that are produced outside the experiment, as well as inside the lead and copper, e.g. through muon interactions (shown as yellow in Fig. 2.1).

- The air of the Gran Sasso laboratories contains ^{222}Rn gas, which appears in the decay chain of ^{238}U. It decays under the emission of α- and β-particles. To protect the detectors from radon, the lead and copper shielding are surrounded by the *radon box*, a gas-tight container that is constantly flushed with radioactively clean nitrogen gas and kept under slight overpressure (shown without color in Fig. 2.1, between muon veto and lead layer).

Having introduced the large hardware structures of the experiment, we will focus in the following section on the detectors themselves and which kind of signal we expect.

2.2 Cryogenic Particle Detector Modules

A CRESST-II or CRESST-III *detector module* consists of a phonon and a light detector, built jointly within a case with scintillating surface on the inside. There are many different designs, from which we will study only that of the TUM40 module, used in CRESST-II [38]. However, most of the principles apply to other detector modules as well. In Fig. 2.2 a schematic drawing of the TUM40 detector module is shown.

As material for the 248 g heavy target crystal (block-shaped, blue in Fig. 2.2) $CaWO_4$ is used, which scintillates visible, blue light. On the lower side of the crystal, a *transition-edge sensor* (TES, Fig. 2.2, red) is mounted. We will explain the working principle of a TES shortly after introducing the hardware of the module. It is a characteristic of the TUM40 module, that the TES is not evaporated onto the target crystal directly, but onto a smaller crystal called the *carrier*. The carrier is then glued to the target. Both crystals are made of the same material. This so-called *composite design* spares the target crystal the high temperature ($\sim 400°$C) that is needed for the evaporation process, as this would compromise its scintillation capability [38]. The target crystal, together with its TES, is called the *phonon detector*. It is another characteristic of the TUM40 module that the target crystal is not held by clamps, but

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

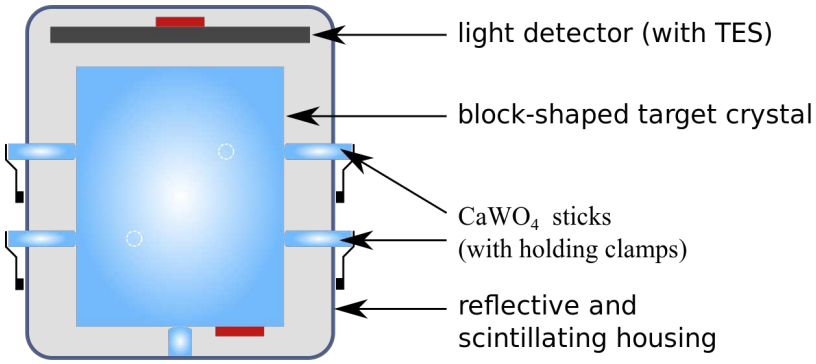


Figure 2.2: Schematic drawing of the TUM40 detector module. The target crystal (block-shaped, blue) together with its TES (red) makes up the phonon detector and works as a highly sensitive thermometer. The material of the target crystal scintillates, the emitted light induces a temperature raise in the light detector (black), also measured with a TES (red). The working principle of the TES is explained in the text body and in Fig. 2.3. The $CaWO_4$ sticks are a characteristic of the TUM40 module and make an identification of particle interactions within the sticks possible [31].

by sticks made out of $CaWO_4$ themselves. That way we can tag particle interactions that occur inside the sticks and remove them in the data analysis. In CRESST-III, crystals are typically by a factor of 10 lighter and therefore more sensible to low-energetic interactions.

The amount of light produced in a particle interaction within the target crystal depends on the energy and the type of the interaction. The light gets reflected and magnified by the inner surface of the housing, which is reflective and scintillating itself. It is then collected by a second detector, a *silicon-on-sapphire (SOS)* disk that has a TES evaporated as well. The SOS-disk together with its TES is called the *light detector*. The light detector mounted inside the TUM40 module is called *Michael* [37]. It was a tradition for several years to give male names to light detectors and female names to phonon detectors. This tradition is today only partially continued – e.g. the phonon detector of the TUM40 module is just called *TUM40*.

The TES of both detectors work the same way. A tungsten film is evaporated on the surface of either the carrier or the SOS disk. The tungsten, which is a superconductor, is cooled to the inflection point from its normal conducting state to its superconducting state. A small increase in the detector's temperature causes therefore a large increase in its resistance. The TES is a highly sensitive thermometer, capable of detecting temperature increases that come from single particle interactions within the detector. The transition curve of a TES is shown in Fig. 2.3. The red line at 18.5 mK marks the temperature at which the detector is operated. The second red line slightly above

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

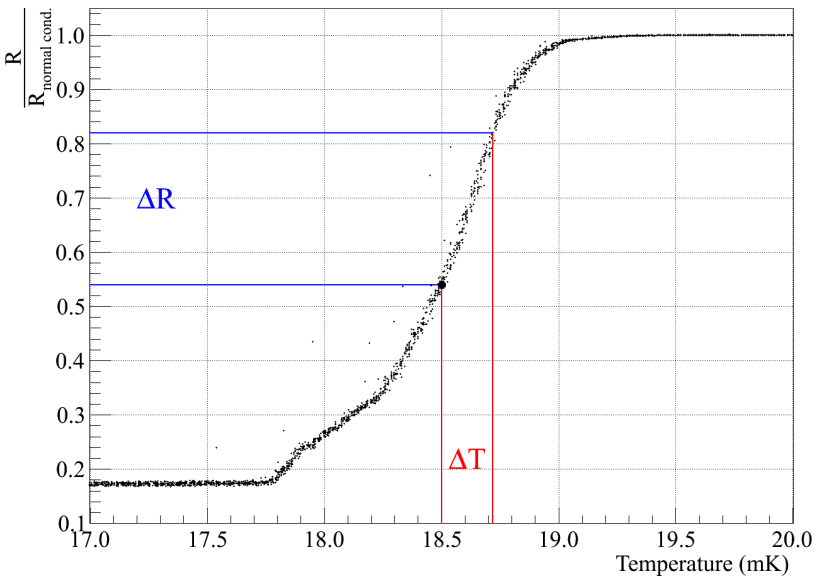


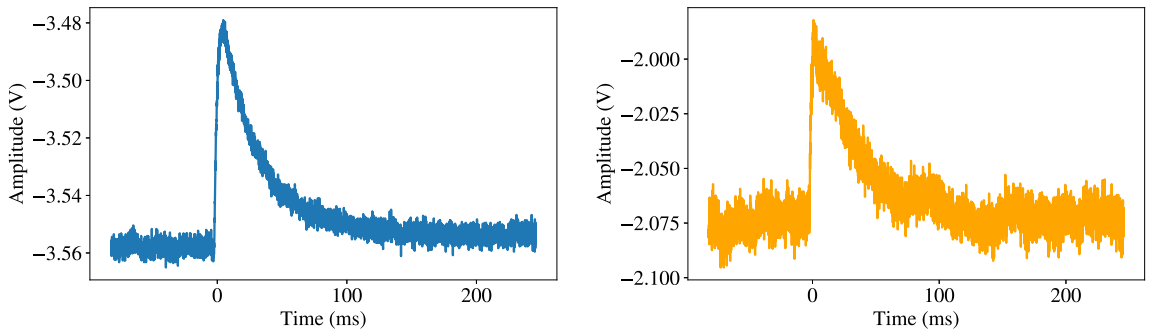
Figure 2.3: Transition curve of the TUM40 transition-edge sensor. The sharp increase in the resistance at a certain temperature is typical for superconducting materials and marks the transition from the normal conducting to the super conducting state. The y-axis shows the resistance, normalized to the resistance in the normal conducting state. A TES that is operated near the inflection point of the transition reacts to small temperature changes (ΔT, red) with a high change in its resistance (ΔR, blue) [31].

18.7 mK, together with the blue lines, show the significant resistance increase of about 50%, caused by a temperature increase of only 2 mK.

The TES is connected in a parallel circuit to a coil and appropriate shunt resistors. The change in resistance in the TES causes a change in the magnetic field that is induced by the coil. The coil is in turn located next to a *Superconducting Quantum Interference Device (SQUID)*, that outputs a voltage signal proportional to the change in the magnetic field. At typical signal form of the phonon channel, originating from a particle interaction in the target crystal, is shown in Fig. 2.4a. The corresponding light signal is shown in Fig. 2.4a. We can see that the signal has the form of a pulse, superimposed to the noise background of the detector and electronics. Events originating from particle interactions always show this characteristic pulse shape. The physics behind the puls is discussed in more detail in chapter 3.

The voltage pulse signals, shown in Fig. 2.4a and 2.4b, scale in a good approximation linearly with the energy of the particle interaction in the target crystal. As mentioned before, the ratio of produced light to heat, i.e. the height of the light and the phonon pulse, depends on the type of the interaction. Electromagnetic recoils typically produce more light than nuclear recoils. Also different types of nuclear recoils are distinguishable. For a meaningful quantification of this relation, we introduce the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) (b)

Figure 2.4: Example of measured voltage pulse traces of (a) the TUM40 phonon detector and (b) the Michael light detector, originating from the same particle interaction in the target crystal. The height of the pulse scales in good approximation linearly with the energy of the particle interaction in the crystal. We can see that the pulse of a light event has a faster rise time and the noise background in the light channel shows a higher amplitude, compared to the phonon channel.

so-called *light-yield parameter*:

$$\text{Light Yield} = \frac{E_l}{E_p}, \tag{2.1}$$

where E_l is the energy measured in the light channel and E_p the energy measured in the phonon channel. The mapping between the measured volt values and the energies of the interactions, defined in the energy calibration, which is briefly discussed in the next section. The light yield is a powerful discriminator and most of the CRESST analysis is based on it. The construction and analysis of the two dimensional light yield vs. phonon energy distribution of all events of one run is the task of the raw data analysis and the high level analysis, respectively. We will explain this in more detail in the following section.

2.3 Analysis of CRESST Data

The data analysis factorizes into two tasks: The first task is the raw data analysis, which processes the raw data and extracts energy and timing information of recoil events. The second task is the high level analysis, which calculates dark matter exclusion limits from the statistics of the recoil events. This thesis focuses on the raw data analysis, however we will also introduce the main concepts of the high level analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

2.3.1 Raw Data Analysis

We introduced in section 2.2 the concept of CRESST detectors and the light yield parameter. In Fig. 2.4a and 2.4b typical phonon and light detector signals are shown. These signals are digitized with a sample frequency of 25 kHz. In CRESST-II events were triggered by a hardware trigger and only triggered events, i.e. ones that exceed a certain threshold, were recorded. The detection threshold, and with that the ability to detect low energetic events, was therefore fixed from the beginning. This was not a problem for the purposes of the experiment in that stage, but it certainly is a limiting factor for the purpose of CRESST-III that is optimized for the detection of low-mass dark matter. Therefore in CRESST-III the continuous data stream is recorded, creating the possibility of defining and optimizing the trigger threshold during the analysis process. An efficient method for the detection of pulse shapes in the voltage pulse trace is the *optimum filter* [39, 40]. It is used in the most current CRESST analysis and makes a trigger threshold of only 30.1 eV possible. The method is explained in detail in chapter 4. For reference, the trigger threshold assigned to the TUM40 module in CRESST-II was 600 eV [36].

The height of the pulse, i.e. the energy information, is distorted by the noise background if we naively evaluate it by taking the highest value of the event. The optimum filter provides a reliable height reconstruction above the trigger threshold. However, for events that exceed the threshold on the phonon channel, but not on the light channel (or vice versa), the energy information of the channel below the threshold would be biased towards too high values. This is problematic for a statistical analysis. We therefore choose another method for the height reconstruction, namely the standard event fit. In this method a typical pulse shape, called *standard event*, is fitted with the raw event. The height of the fitted standard event is taken to be the height of the pulse. This method is investigated in detail in chapter 4. An important feature of the standard event fit is that it might yield negative results for the pulse height, especially for very small pulses. This is a desirable property as it makes the unbiased determination of the average energy of low-energetic events possible.

A CRESST analysis is always based on the data of an entire a whole *run*. A run is the time period of data taking with the same detector setup, typically lasting between 6 and 12 months. The TUM40 module was operated in run 33, which took place in CRESST-II phase 2. Its total exposure amounts to 29 kg days. At the time of writing run 36 is in preparation, while results of run 34 were published in 2019 and run 35 currently is analyzed. The process of data taking gets usually interrupted by the refills of liquid helium into the tanks above the cryostat. Such interventions are necessary

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

three times a week. The refill causes vibrations that destabilize the temperature of the experiment, which renders a reliable reconstruction of the energy of events impossible for several hours. The stability of the *operating point*, i.e. the temperature of the detectors, is controlled with *control pulses* that are injected every 3 to 6 seconds [37]. These are voltage pulses sent to a heater that is mounted either on the detector crystal or directly on the TES. The control pulses have a similar shape as particle pulses (see Fig. 2.4) have, but they are large enough to drive the thermometer out of the superconducting phase, i.e. they saturate the TES. The resistance of the TES in the normal conducting state is a constant (see limit on the right side of the transition curve in Fig. 2.3). Therefore the height of the saturated control pulse provides information about the distance of the operating point to the top of the superconducting transition, i.e. about the current temperature of the TES. This information is exact on the scale of $\mathcal{O}(\mu\mathrm{K})$ and is used to control a steady heating current, that keeps the operating point stable.

As mentioned, a stable operating point is important for a reliable energy reconstruction of the events. At the start of a new run the energy scale is calibrated with a ^{57}Co source, placed outside the shielding. The source emits γ-particles at 122.1 keV and 136.5 keV. These sources provide fixed points on the measured energy scale. The heights of the voltage pulses of CRESST-II detectors scale up to these energies almost linearly, therefore these fixed points can be used to define a linear map, i.e. a scalar conversion factor, between voltage pulse traces of the phonon and light channel and the energy of the particle interaction. This factor is called the *energy-voltage conversion (EPC) rate* and for the TUM40 phonon detector it has a value of (111.24 ± 0.06) keV/V, for the corresponding light detector Michael (67.92 ± 0.07) keV/V. To provide a precise energy calibration, even months after the start of the run and numerous refill processes, *test pulses* are injected into the heater of the detector module roughly every 30 seconds. These test pulses are similar to the control pulses, but with amplitudes in the same range as those of recoil events. The deviation of the height of a test pulse from those with the same injected voltage at the time of energy calibration allows for a correction of small time-dependent drifts of the detector response.

For CRESST-III, the lines of a cobalt source above 100 keV exceed the linear range of the TES and lead to saturation. This imposes a tremendous problem for the calibration of some detector modules, rendering some of the recorded data of run 34 and 35 useless. However, some of the data was made accessible for the energy calibration by an ingenious method, using a truncated fit systematic [41]. The truncated fit is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

explained in detail in chapter 3. For other detector modules, the cosmic activation lines originating from ^{182}W were used for the energy calibration instead.

The last step in the raw data analysis is the filtering of events that do not originate from particle recoils or whose energies cannot be reconstructed reliably. Such selections are called *cuts*, over which we give a qualitative overview below:

- The *rate cut* filters time domains in which abnormally many events occur. These are most probably caused by some external source, that came close to the setup of the experiment at the given time, e.g. a moving crane of the underground laboratory.

- The *stability cut* removes time domains in which the height of the test pulses deviate strongly from the mean value of the test pulses with the same injected voltage value.

- *Coincidence cuts* remove events that coincide with the muon veto, a signal in the sticks or other detectors. These cannot originate from a weak interaction in one of the detectors, as the recoil energy of such an interaction would not cause any signal in the other detectors.

- *Quality cuts* are meant to remove events that do not show the sought-for pulse shape. To accomplish this, various different techniques are applied. In chapter 5 we will introduce the carrier event, as a well-studied example of such an artifact. It is caused by a recoil event that happens inside the carrier. Other causes for anomalous events are voltage spikes, squid resets, noise triggers, flux quantum losses, pile-ups, strong baseline tilts or light only events – depending on the detector design, there are more. Ref. [37] provides a good explanation of most of them. Most quality cuts are defined and optimized using fitting techniques, comparable to the above mentioned standard event fit. In chapter 5 we will also study the standard event fit for the detection of anomalous events.

We test the quality of cuts by calculating the survival rates of simulated recoil events, called *efficiencies*. Simulated events are standard events of desired height, that we superimpose with a noise baseline. The simulation process is explained in detail in chapter 3. Using simulated instead of measured events has two advantages: On the one hand we can be sure that no anomalous events occur when we test our cuts. On the other hand we can control the energy of the simulated events and therefore quickly construct a sufficiently large set of simulated pulses, to reach statistically significant

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

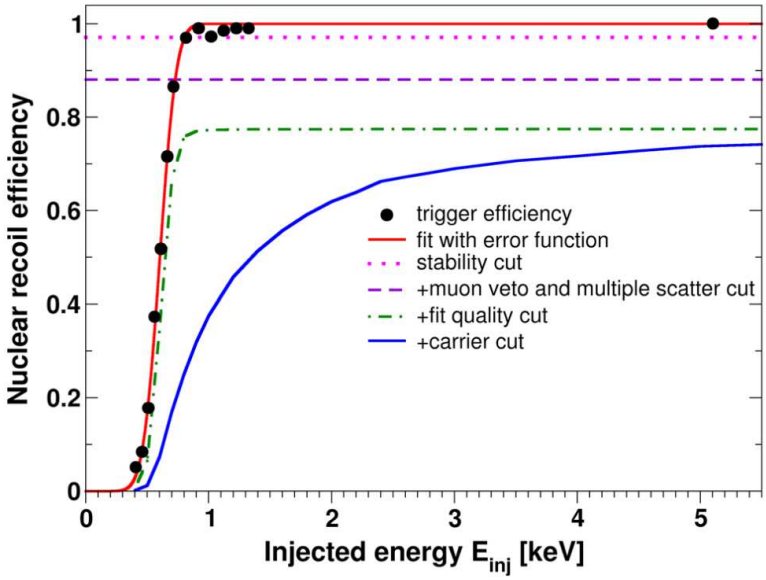


Figure 2.5: Nuclear recoil efficiencies for the TUM40 module (run 33). The efficiency is the survival rate of simulated nuclear recoil events that pass through the analysis chain. The filled, black circles are trigger efficiencies measured by injecting heater pulses with closely spaced discrete energies. The full (red) curve is a fit with an error function which yields an energy threshold of (603 ± 2) eV. Also shown in this plot are the nuclear recoil efficiencies after application of the stability cut (pink, dotted), the coincidence cuts (purple, dashed) and the quality cuts (green, dash-dotted). The rate and coincidence cuts do not depend on the energy of the event, because they remove a certain time interval and are not based on the shape of removed event itself. The efficiency after the carrier cut is shown separately (blue). Figure taken from Ref. [36].

results with our efficiency calculation. In Fig. 2.5 an example of calculated efficiencies for the analysis chain of the TUM40, that was used in CRESST-II run 33, is shown before and after all cuts. The trigger threshold for the TUM40 detector was chosen to be 0.6 keV. The fitted error curve (Fig. 2.5, red) confirms this threshold.

It is not common in practice to calculate the survival rate of anomalous events. Such an anomalous event that survives the raw data analysis is counted as recoil event. As all calculations of exclusion limits are done conservatively, every recoil event with appropriate light yield and phonon energy is taken as a dark matter candidate event and weakens the exclusion limit. It might be helpful to perform a cross check by calculating survival rates of anomalous events to fine-tune quality cuts. This is however only efficiently possible for artifacts that are well enough studied to provide simulations of them, as we did it for the efficiency calculations. By now, only carrier events are well enough studied for such tests. We will investigate these in chapter 6.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

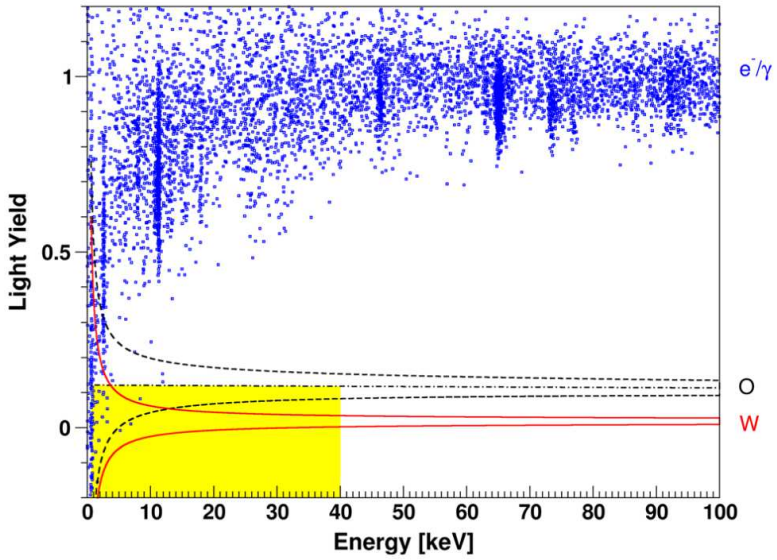


Figure 2.6: Light yield plot of the TUM40 detector module (run 33). The light yield parameter is defined in Eq. 2.1, the energy on the abscissa is the energy of the phonon channel. Every blue point corresponds to an event. The band-like accumulation of events with a light yield parameter of roughly 1 are the e^-/γ-events. Gaussian band fits are done on the oxygen-recoil events (black) and the tungsten-recoil events (red). The continuous lines correspond to the intervals that include 80% of the events of the corresponding band. All events inside the acceptance region (yellow) are counted as candidate dark matter events in the calculation of exclusion limits [36].

To summarize, the raw data analysis consists of four steps: Hard- or software triggering of events, reconstruction of the pulse heights, reconstruction of the event energies and filtering of anomalous events. At the end of the raw data analysis chain we have extracted the energy and time information of every particle event. This information is the basis for the high level analysis, over which we will give a compact overview in the following section.

2.3.2 High Level Analysis

The result of the raw data analysis, that we introduced in the previous chapter, is the two dimensional distribution of events, depending on the light yield parameter and the reconstructed energy of the phonon channel. We call this distribution the *light-yield plot*. An example of a light-yield plot, calculated with data from the Lise detector module from CRESST-II run 33, is shown in Fig. 2.6. The light-yield parameter has strong discriminating power over the different types of events. As the energy calibration was done on e^--events, electromagnetic recoils (including γ-events) have a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

light-yield parameter of roughly one and separate well from the nuclear recoil events. They build a band-like form in the light-yield plot, same do different types of nuclear recoil events. A Gaussian band fit is calculated on two bands of nuclear recoils, that can be identified in the light-yield plot: The oxygen-nuclei recoil band (Fig. 2.6, red) and the tungsten-nuclei recoil band (Fig. 2.6, green). The yellow region marks the region of interest, i.e. events in this region that survive the raw data analysis chain are counted as candidate dark matter events.

The Yellin method is a common method for the calculation of exclusion limits [42]. However, the CRESST collaboration developed their own method, based on a likelihood analysis. We will not explain the process of calculating of dark matter exclusion limits in this thesis, a detailed explanation of both methods can be found in Ref. [43].

In this chapter we introduced the setup of the CRESST experiment and the working principle of its detector modules. We gave an overview over the raw data and high level data analysis. In the following chapter we will study the voltage pulse traces in detail and introduce a physical motivation of their form and an efficient simulation method for arbitrarily large sets of data. With those simulated data sets, we will in the chapters 4 to 6 investigate machine learning methods for the raw data analysis, namely for the tasks of triggering events, reconstructing height information and filtering carrier events.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Chapter 3

Data Simulation

Most data analysis methods used by the CRESST collaboration are *knowledge-based*, i.e. based on physics knowledge about the data. One example is the standard event fit, where we fit a typical pulse shape on raw data events. If the event is shaped like a nominal pulse itself, the residuals of the fit are lower than otherwise. All raw data events above a defined error threshold are suspected to be non-recoil events and are therefore discarded. The pulse shape is a type of knowledge we provide to our method, as well as the error threshold. Different to those methods, *machine learning* is not or only partially based on physics knowledge. It learns from provided data and is practical for specific tasks in the analysis of high dimensional and highly correlated data that do not display an easily comprehensible structure. However, not every method is clearly knowledge-based or a machine learning method. There are many hybrids, e.g. knowledge-based cuts with parameters optimized by machine learning.

Machine learning is rather new to the CRESST analysis, therefore most of our data has yet to be prepared for these methods. This thesis covers two learning types: *Supervised learning* and *unsupervised learning*. Supervised learning is a statistical approximation method. It needs a data set, each element of the data set consists of a *sample* x and a *label* y. A supervised method learns in a process called *training* to approximate a function $\hat{f}$ that correctly assigns a label to a given sample:

$$\hat{f}(x_i) \approx y_i. \tag{3.1}$$

Both data and sample can either be categorical or numerical, i.e. it can be a *regression* or a *classification* problem. A prominent example for supervised classification is image recognition: A supervised learning technique can learn to recognize faces or objects on pictures, provided a large and diverse data set of labeled pictures with faces or similar objects. Many supervised learning methods work incredibly well

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

today [44–51]. However, their limitation is the need for labeled data, which is not available for every application. We will study machine learning regression methods in chapter 4 and classification methods in chapter 6. At this point we have a remark on our nomenclature: It is common in literature, to call the elements of the data set samples and labels. However, it is also common to call a single data points of a time series sample, and their recording rate sample frequency. To make to situation even worse, we sometimes refer to a small subset of a set as sample, which is the common name in probability theory. We hope that this will not lead to confusion and do our best to state clearly which kind of sample is meant.

Unsupervised learning does not have this limitation, the elements of a data set for unsupervised learning consist only of samples without labels. However, unsupervised methods do usually not perform very well for classification or regression problems, they rather search for yet unknown structure in the data. Typical use cases are clustering algorithms and anomaly detection. We will study unsupervised methods for anomaly detection and visualization in chapter 5.

In this thesis, we mainly use *deep learning*, i.e. machine learning with neural networks. In appendix A we provide an introduction to the used deep learning methods.

All machine learning methods have the need for cleaned and prepared data in common, in the supervised case also labels for all samples are needed. As the CRESST data is, same as most data from physics experiments, initially neither cleaned nor labeled, we follow for this thesis a different approach in order to provide a data set for supervised and unsupervised learning methods: We simulate data that resembles the features of the TUM40 detector module. One advantage of simulated data is the ability to create arbitrarily large data sets and exact knowledge of all physics features of each sample. The second advantage is of practical nature: The simulation process only takes several minutes, we therefore do not need to store large data sets prepared for the training of machine learning models but can save the simulation code and all relevant parameters instead.

This chapter introduces our methods for the simulation of machine learning data. In the first section the pulse shape model is derived, an analytical function that describes the typical pulse shape of recoil events very well. After that, we present a method to simulate realistic noise baselines and superpose them with the pulse shape template to obtain simulated recoil events. The chapter concludes with a short study of saturated pulses and the attempt to reconstruct their energy information.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

3.1 The Pulse Shape Model

We are in the merry position of knowing a physical model to approximate the response of our detectors. It is based on a thermal equivalent circuit [52] and describes our data very well. We only introduce a simplified version of the model derived in Ref. [7], which yields the same results for our usage. The principles apply to every CRESST detector, i.e. to phonon and light detectors alike.

The temperature of the detector crystal is given by the temperature of its phonons T_c, while the temperature in the thermometer, which is a TES in our case, is mainly determined by the temperature of the free electrons T_e. Both systems have an individual heat capacity and are thermally linked with a thermal resistance G_{ec}. The housing, which is linked with a copper wire to the cold finger of the cryostat, serves as a heat bath with a constant temperature T_b. We denote the thermal resistance of the crystal phonons and the thermometer electrons to the heat bath with G_{cb} and G_{eb} respectively. Heat outputs $P_c(t)$ and $P_e(t)$ are transferred to the crystal phonons and the thermometer electrons by an external source - this can either be a particle interaction within the crystal or the heater that is attached to the TES and used for fine tuning of the temperature. The situation is visualized in Fig. 3.1.

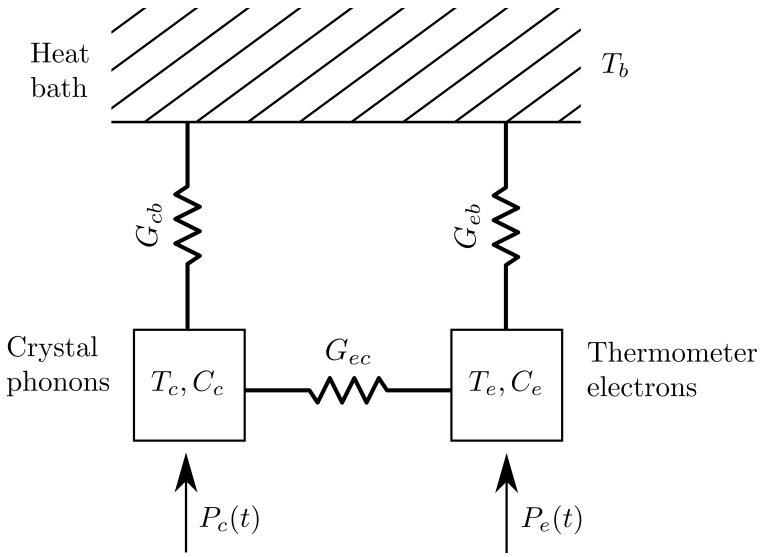


Figure 3.1: Thermal circuit of the crystal phonons with temperature T_c, the thermometer electrons with temperature T_e and the heat bath with temperature T_b. G_{ec}, G_{cb} and G_{eb} are the thermal resistances between the thermometer electrons and the crystal phonons, the crystal phonons and the heat bath as well as between the thermometer electrons and the heat bath. Heat outputs $P_c(t)$ and $P_e(t)$ are transferred to the crystal phonons and the thermometer electrons. Figure based on Ref. [7].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

We can describe the temperature of the crystal phonons and thermometer electrons with the differential equations

$$C_c \frac{dT_c}{dt} + G_{ec}(T_c - T_e) + G_{cb}(T_c - T_b) = P_c(t), \tag{3.2}$$

$$C_e \frac{dT_e}{dt} - G_{ec}(T_c - T_e) + G_{eb}(T_e - T_b) = P_e(t). \tag{3.3}$$

A particle interaction in the detector crystal creates an energy deposition ΔE in the crystal lattice. The freed energy manifests itself as non-thermal phonons that decay exponentially on a time scale τ_n at the surface of the crystal as well as in the thermometer film directly. The share of non-thermal phonons decaying in the thermometer film is denoted by ϵ. The heat output from the particle interaction to the crystal phonons and the thermometer electrons can be written with the equations

$$P_c(t) = \theta(t)\epsilon P_0 e^{-t/\tau_n}, \tag{3.4}$$

$$P_e(t) = \theta(t)(1-\epsilon) P_0 e^{-t/\tau_n}, \tag{3.5}$$

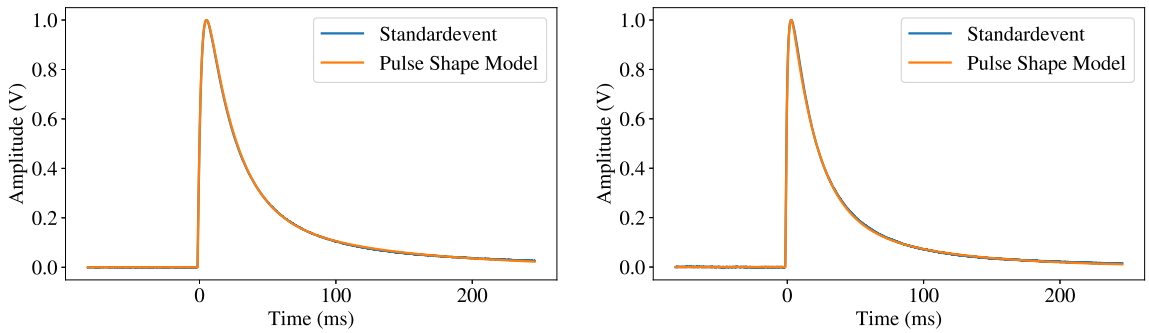
where P_0 is a normalization constant fixed by the total energy deposition and $\theta(t)$ is the Heaviside step function, i.e. a function that equals zero for $t < 0$ and one for $t \geq 0$. We solve the coupled system of differential equations (3.2), (3.3) for the temperature difference between the thermometer electrons and the heat bath $\Delta T_e(t) := T_e(t) - T_b$. Solutions for the inhomogeneities (3.4), (3.5) have the form

$$\Delta T_e(t) = \theta(t) \left[A_n (e^{-t/\tau_n} - e^{-t/\tau_{in}}) + A_t (e^{-t/\tau_t} - e^{-t/\tau_n}) \right], \tag{3.6}$$

where τ_{in} and τ_t are parameters describing the internal thermal relaxation time of the thermometer and the whole system. A_n and A_t are the amplitude of two distinct components of the temperature pulse. All parameters are determined by the heat capacities, heat resistances and time scales within the system. An important property is, that the amplites scale linearly proportional to the deposited energy of the recoil, i.e. $A_n, A_t \sim \Delta E$. However, the exact equations for all paramters are rather complicated and do not contribute to the purpose of this thesis. They can be found in Ref. [7], together with a detailed derivation of the solution (3.6).

We can see that the particle interaction introduces a temperature pulse with two components for the thermometer electrons. The first component origins from the decay of the non-thermal phonons directly in the thermometer and is called the *non-thermal component*. The second component, the *thermal component*, origins from the temperature rise of the crystal that is caused by the decay of non-thermal phonons on its surface.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) (b)

Figure 3.2: Standardevent (blue) and fitted pulse shape model (orange) for (a) the TUM40 phonon detector and (b) the Michael light detector. The fit function matches the standard event very well, indicating a good model of our data. The standard event is scaled to a height of one Volt.

To verify the validity of function (3.6) we perform a fit on a *standard event*. A standard event is a typical recoil event with boosted signal-to-noise ratio, which is achieved by superposing several hundred pulses with similar height. The fit is done by minimizing the mean squared error between function (3.6) and the standard event with respect to the five parameters $A_n, A_t, \tau_n, \tau_{in}$ and τ_t of the function and an additional time shift t_0, accounting for a possible discrepancy between the onset of the pulse and the origin of the time scale. The results for a TUM40 (phonon) and a Michael (light) recoil event are shown in Fig. 3.2. The fit function (orange) matches the standard event (blue) very well, verifying equation (3.6) as a good model for our data.

We can distinguish between two different regimes in the operation of cryogenic detectors. The *bolometric mode* (energy flux mode) is the case where the thermometer relaxation is much faster than the non-thermal photon decay rate, i.e. $\tau_{in} \ll \tau_n$. The thermometer signal is proportional to the power delivered by the non-thermal phonons. In this operation regime, the decay time of the non-thermal component and the rise time of the thermal component are τ_n. The second mode is the *calorimetric mode* (energy integral mode), where the thermometer collects non-thermal phonons faster than they can relax into the heat bath, i.e. $\tau_{in} \gg \tau_n$. The output of the thermometer is therefore proportional to the deposited energy and both components of the pulse have a rise time of τ_n.

In CRESST-II all phonon detectors are operated as bolometers and all light detectors as calorimeters. The smaller detectors in CRESST-III are all operated in the calorimetric regime.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

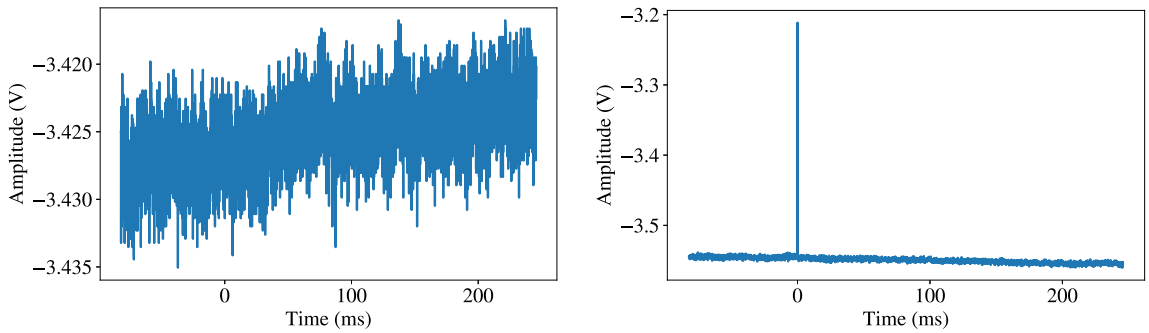
For now, we ignored the *self-heating* effect of the TES that influences the temperature rise and fall and therefore the pulse shape of all detectors. We will study it in section 3.4, together with the saturation of high energetic pulses. Before that, we introduce a method for the simulation of detector noise in the following section.

3.2 Simulated Noise Baselines

In CRESST-II all data recordings are done with a readout window that has a length of 8192 samples, which equals 328 ms. The raw pulse traces of three types of events are recorded by the data acquisition: Injected *test pulses*, i.e. heater pulses for energy calibration, triggered detector events and empty noise baselines. For the simulation of recoil events, we want to superpose empty noise baselines with templates from our pulse shape mode. However, not all recorded empty noise baselines are actually empty. Roughly 2% of all recorded baselines accidentally contain artifacts or pulses. This happens due to the algorithm of the trigger and data acquisition: The trigger is blocked while a recording is made. Therefore when an event is triggered the data acquisition stores it in a buffer to wait for additional triggers on other channels. If the impulse for recording an empty noise baseline accidentally coincides with an non-empty recording buffer, the triggered event is recorded instead of an empty baseline. An example of a recorded noise baseline from the TUM40 detector is shown in Fig. (3.3a). A baseline with a delta voltage spike, originating in the electronics of the data acquisition, is shown in Fig. (3.3b). The buffer of the trigger algorithm serves a second cause: For every triggered event, a pre-trigger time of 80 ms is recorded, such that the triggered event can be placed in the second quarter of the recorded window. By that we obtain an information about the baseline offset and drift at the time of triggering.

Noise baselines contain information about the *noise power spectrum (NPS)* of the Gaussian detector noise and a possible drift structure of the baseline. The noise power spectrum is defined as the squared absolute value of the Fourier coefficients of empty noise baselines. We usually average the NPS of several hundred noise baselines to distinguish between typical features and randomly occurring fluctuations in the spectrum. The spectra of the TUM40 phonon detector and the Michael light detector are shown in Fig. 3.4 in blue and orange, respectively. We can observe characteristic high frequencies around 50 Hz, which is due to the mains voltage. The second harmonic at 150 Hz is visible, as well as some typical peaks above 10 kHz. These are probably caused by the operation frequencies of electrical components. Spikes and other high-voltage artifacts strongly distort the noise power spectrum of the empty

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) (b)

Figure 3.3: (a) Recorded empty baseline from the TUM40 detector, used in run 33. We can see a slight drift in the baseline, superimposed with Gaussian noise. (b) A delta voltage spike, originating from the electronics of the data acquisition system is accidentally recorded instead of an empty baseline. About 2% of all empty baselines contain such artifacts.

baselines by contributing high amplitudes of frequencies above 100 Hz. The mean power spectrum of triggered events of the detectors are also shown in Fig. 3.4 (green: TUM40, red: Michael). We can see, that the events have a power spectrum with overall higher amplitude and especially stronger amplitudes for frequencies between 100 Hz and 1000 Hz than the noise. We want to train supervised learning models on the simulated raw data. For such training processes a few malicious or wrongly classified samples in the data set can worsen the performance of the trained model significantly. To avoid this, we also simulate the noise baselines and by that take total control over all features of the samples in our data set. A comparison of the differences between simulated and measured noise baselines was also done in Ref. [14].

A method for the simulation of noise baselines is studied in Ref. [53] and was applied to CRESST in Ref. [41]. This method yields baselines of arbitrary length and with a given noise power spectrum. However, it does not consider a possible drift structure in the baseline. For most use cases the drift structure of the noise baseline is small compared to the size of the signal and therefore negligible. In CRESST, we are especially interested in low energetic events, i.e. in small pulses. These low energy events typically show pulse heights in the same order of magnitude as the size of the baseline drift. Our method builds upon the method presented in Ref. [53], but also includes the information of the baseline drifts. We describe the algorithm in the following step by step.

As a first step we prepare a small data set of the numerical arrays of several hundred recorded noise baselines $b_i[x]$. We denote the array index with x to avoid

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

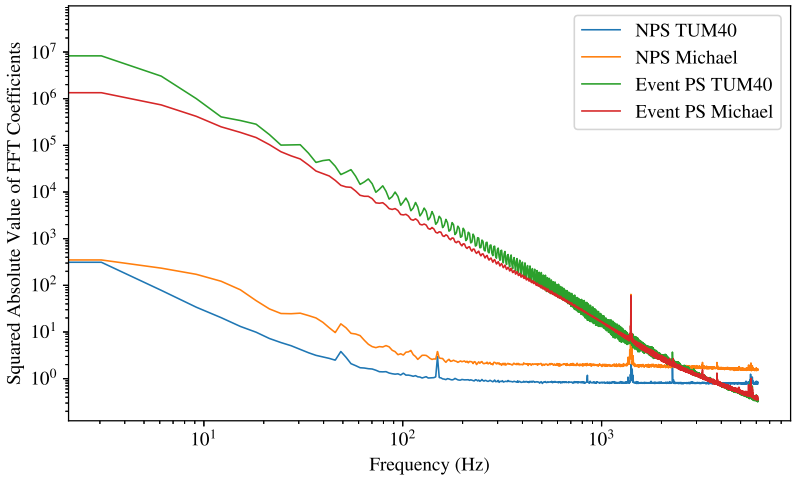


Figure 3.4: Noise power spectra of the TUM40 phonon (blue) and the Michael light detector (orange) and mean power spectra of triggered events on the TUM40 phonon (green) and Michael light (red) detector. We can see that the noise background of the light detector has an overall higher amplitude. The prominent peak in the noise power spectra at 50 Hz is due to the mains electricity.

confusion with the time t and keep the letters $i, j, \ldots$ for sum indices. To extract the meaningful drift information, we need to remove the offset from all baselines, that is caused by the SQUID-readout system and does not contain any useful information, i.e. we subtract from every baseline the mean of its values:

$$b_i[x] \leftarrow b_i[x] - \langle b_i[x] \rangle_x. \tag{3.7}$$

We denote the mean value of an array with respect to the variable x by angel brackets $\langle \cdot \rangle_x$. The arrow pointing to the left is a convenient notation for writing algorithms, as we can overwrite variables with a new value several times. The next step is the extraction of the drift information. We do this by a cubic polynomial mean square error fit, i.e. we find for every baseline b_i the polynomial coefficients $a_0, ..., a_3$ that minimize the error functional:

$$\{a_0, a_1, a_2, a_3\} \leftarrow \underset{a_0,a_1,a_2,a_3}{\arg\min} \sum_x \left(b_i[x] - \sum_{j=0}^{3} a_j t^j[x] \right)^2 . \tag{3.8}$$

As explained above, some of the baselines show artifacts that distort the NPS of empty baselines. Those can typically not be approximated by polynomials very well and show therefore a higher error in the fit than empty baselines. We clean our set of baselines by disposing all baselines whose fit error exceeds the mean error of all baselines by more than one standard deviation. We repeat the fit procedure on the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

cleaned baselines and obtain the polynomial fit coefficient distribution of the cleaned baselines. The coefficients happen to follow in good approximation a multivariate Gauss distribution in their four dimensional parameter space. By calculating their mean vector μ_c and covariance matrix Σ_c we construct the parameters of a distribution from which we can sample new, realistic baseline drifts $\tilde{d}_i[x]$:

$$\{a_0, \ldots, a_3\} \sim \mathcal{N}(\mu_c, \Sigma_c), \tag{3.9}$$

$$\tilde{d}_i[x] \leftarrow \sum_{j=0}^{3} a_j t^j[x]. \tag{3.10}$$

We managed to extract the drift information of the baselines and go on to extract the information of the noise power spectrum.

The mean noise power spectrum of the cleaned baselines $\mathcal{N}[\omega]$ is defined by

$$\mathcal{N}[\omega] \leftarrow \langle \left|\mathcal{F}(b_i[x])\right|^2 \rangle_i, \tag{3.11}$$

where $\mathcal{F}(\cdot)$ is the Fourier transformation and ω a different array index to avoid confusion of functions defined in time and frequency space. For a given noise power spectrum, we can always generate a noise baseline by taking its square root, adding randomly chosen phases and applying the inverse Fourier transformation $\mathcal{F}^{-1}(\cdot)$. We will apply this concept shortly, before that we also need to compute the noise power spectrum of every polynomial drift that we created:

$$\mathcal{N}_i[\omega] \leftarrow \left|\mathcal{F}(\tilde{d}_i[x])\right|^2, \tag{3.12}$$

and then subtract it from the mean noise power spectrum of the cleaned baselines:

$$\mathcal{N}_i[\omega] \leftarrow \mathcal{N}[\omega] - \mathcal{N}_i[\omega]. \tag{3.13}$$

We need to add a noise baseline $\tilde{b}_i$ with the noise power spectrum $N_i[\omega]$ to the sampled drifts $\tilde{d}_i$ to obtain a noise baseline with the desired drift structure and noise. We obtain a baseline like this by taking the root of the noise power spectrum, adding randomly chosen phases to $N_i[\omega]$ and applying an inverse Fourier transformation:

$$\phi[\omega] \sim \mathcal{U}([0, 2\pi]), \tag{3.14}$$

$$\tilde{b}_i \leftarrow \mathcal{F}^{-1}\left(\sqrt{\mathcal{N}_i[\omega]} e^{i\phi[\omega]}\right). \tag{3.15}$$

Here $\mathcal{U}([0, 2\pi])$ denotes a uniform distribution over die interval $[0, 2\pi]$ from that we sample the phases. Finally, we add these noise baselines to the generated drifts

$$\tilde{b}_i \leftarrow \tilde{d}_i + \tilde{b}_i, \tag{3.16}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and, as a result obtain noise baselines with the desired drift structure and noise power spectrum. One example of a simulated baseline and its polynomial drift is shown in Fig. 3.5 in green and orange. We can see a good match of the simulation with the measured baseline depicted in Fig. 3.3a.

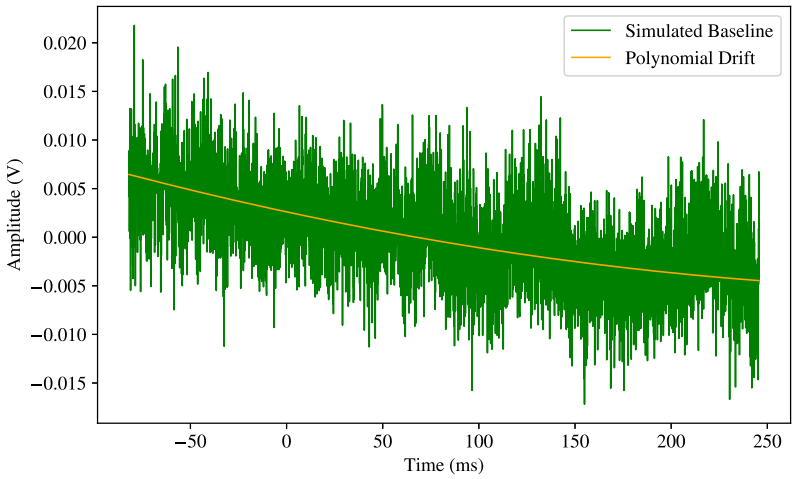


Figure 3.5: Simulated noise baseline (green) with noise power spectrum and polynomial drift structure (orange) of the TUM40 phonon detector. We can see a good match of the simulation with the measured baseline, depicted in Fig. 3.3a.

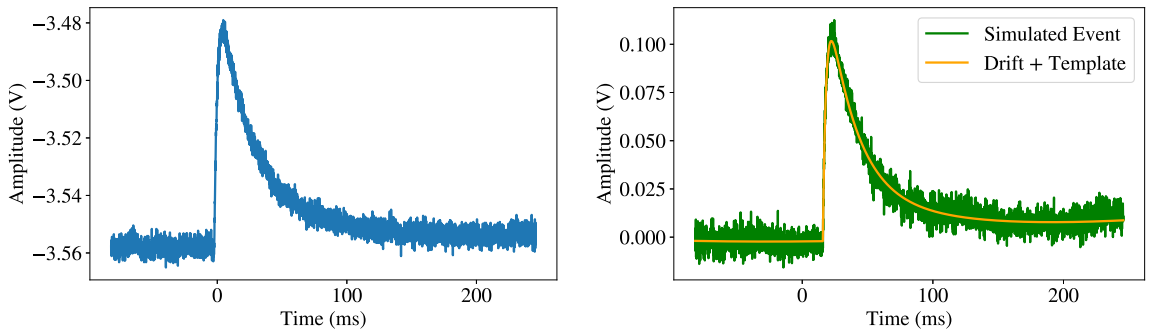
There is also a machine learning based method to simulate noise baselines. It uses generative adversarial networks (GANs) and produces reasonable well results, which are shown in appendix B. However, the knowledge-based simulation method we presented above outperforms the machine learning method and is hence applied for the simulation of the data sets for this thesis.

In the past two sections, we simulated the temperature pulse of particle recoils and the noise background of the detector. The last step is their combination to obtain simulated particle recoil events. This is what we do in the following section.

3.3 The Simulated Recoil Data Set

The temperature of CRESST detectors is calibrated such that the TES operates in the steep, and in good approximation linear, part of its transition curve. Therefore, not only a temperature rise in the thermometer electrons is directly proportional to the voltage rise in the SQUID output, but it is also valid to naively superpose detector events. We do so with our simulated TUM40 noise baselines (see section 3.2) and the pulse shape model, fitted to the TUM40 standard event (see section 3.1).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub



(a) (b)

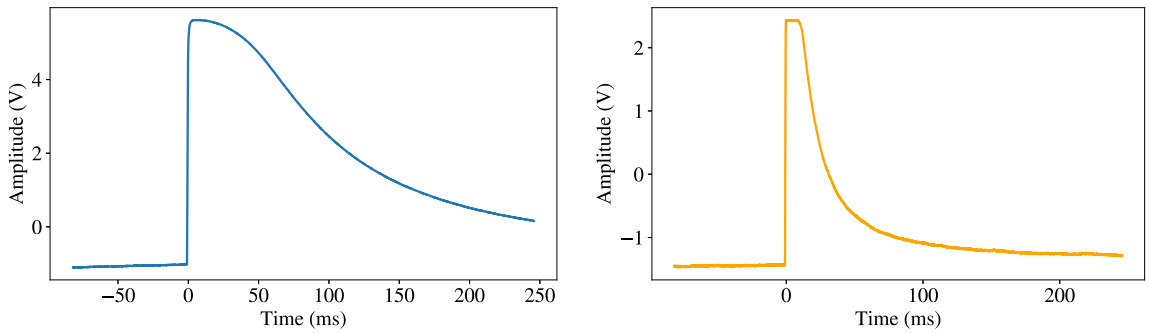
Figure 3.6: (a) Recorded TUM40 recoil event. (b) Simulated recoil event with the TUM40 pulse shape model and noise power spectrum (green) and its baseline drift structure superposed with the pulse shape model (orange).

Additionally, we introduce a time shift, sampled from a uniform distribution over the interval $[-20\text{ ms}, 20\text{ ms}]$, to the pulse shape model. This accounts for possible early or late pulse onsets, that might occur in the recorded events due to the so-called *trigger-walk* effect, a dependency of the trigger time on the pulse energy. One example of a simulated event is shown in Fig. 3.6b (green), together with the drift structure of its baseline added to the pulse shape model (orange). For comparison, a measured recoil event is depicted in Fig. 3.6a, demonstrating the quality of the simulation. We can now simulate recoil events of arbitrary energy, as long as we stay in the linear regime of the TES transition curve.

For our pulse shape model we made the assumption that all temperature pulses caused by regular recoils within the crystal have the same form and vary only in height, proportional to the energy of the recoil. This, however, is probably a simplification of reality for our experiment. The parameters of the pulse shape model depend, inter alia, on the heat capacities of the crystal and the TES and their thermal conductance towards each other and the heat bath. These quantities are temperature dependent themselves and might therefore change with different recoil energies. It would be interesting to fit the pulse shape model not only to the standard event, but to many recoil events and to construct a distribution in their parameter space – same as we did for the polynomial baseline fit coefficients. However, this a matter of ongoing research and not included in this thesis.

So far we excluded the self-heating of the TES from our considerations, even tough it has an effect to the pulse shape in every temperature regime of the TES, i.e. also in the linear range. This effect is omnipresent in any electrical circuit that contains

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub



(a) (b)

Figure 3.7: Saturated recoil events of (a) the TUM40 phonon detector and (b) the Michael light detector. We can see the strong flattened peak, where the TES reaches its normal conducting phase, resulting in an approximately constant resistance for small temperature changes.

components with non-negligible temperature-dependent electrical resistance, which applies to the resistance of the TES. Also, we did only consider the linear range of the TES transition curve, which excludes a description of high energetic events, e.g. originating from α-decays. Both these effects are not the main objective of this thesis, but we introduce them qualitatively in the following, last section of this chapter.

3.4 High Energy Recoils

The transition curve of the TES is in its steep part in good approximation linear. For the TUM40 phonon detector operated in run 33, this linear range holds for voltage pulses up to 1 V. Events with higher energies saturate the TES, i.e. they drive the superconductor in the normal conducting state. In the normal conducting state, the resistance of the TES does barely change for small temperature changes, resulting in a flattened peak of the voltage pulse. In Fig. 3.7a and 3.7b we show two strongly saturated phonon and light events of the TUM40 detector module, both reaching the normal conducting state of the TES at their peak.

This problem is usually fixed with the *truncated fit* [37]. This is a fit of only the low energetic part of the pulse, i.e. the front part of its rising edge and the rear part of its tail. These are the parts of the pulse that are not saturated and should therefore match the form of the standard event. The samples in between are excluded from the minimization functional corresponding to the fit. After the fit is performed, we can reconstruct the unsaturated height of the pulse by adding the peak of the non-truncated fit template to its truncated version. The situation is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

depicted in Fig. 3.8. The truncated template (orange) fits the saturated event (blue) in its non-saturated parts. The saturated part is excluded from the fit, visualized with an orange dashed line. The blue dashed line shows the reconstructed peak of the event, also reconstructing its energy information. This method was used for the energy calibration of some detectors, as the recoil energy of events originating from the ^{57}Co calibration source (> 120 keV) exceeds the linear range for most CRESST-III detectors [41].

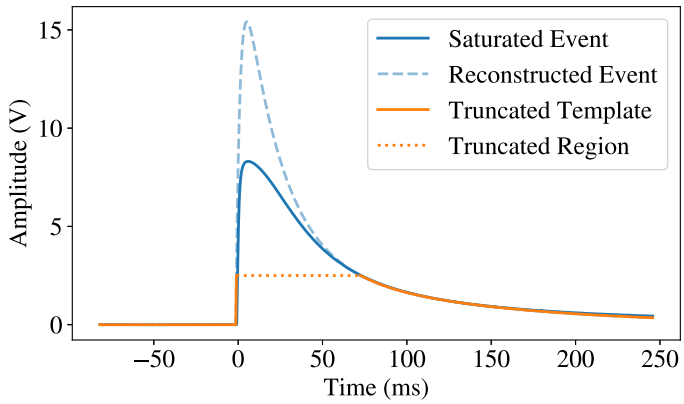


Figure 3.8: Illustration of the truncated fit. A saturated TUM40 phonon pulse (blue) is fitted with a truncated template (orange) in the non-saturated regime. The saturated, or truncated, samples (orange, dotted) are excluded from the fit. The height of the pulse is then reconstructed by adding the non-saturated peak to the template (blue, dashed).

The truncated fit method is work well for many application. However, recent observations on test pulses indicate a possible weak stop of this method. When we perform the truncated fit, we make the implicit assumption that the pulse shape does only saturate on this peak, i.e. in its height. We assume that there is no saturation in its width, which seems plausible because the superconductor does not feature any type of hysteresis on the upper edge of its transition curve. If the pulses would also saturate in their width, we would reconstruct energies with the truncated fit that feature a bias towards lower values than the true values of the energies.

We tested this assumption on the test pulses that were injected into the heater of the TUM40 phonon detector and the Michael light detector in run 33. These had, in the non-linear regime, voltage values of 1 V, 2 V, 3 V and 10 V. It might be contra-intuitive, that we expect the pulse height to scale linearly with the injected voltage values in any range, because the power that turns into heat within the heater scales quadratically with the injected voltage. However, the heater electronics automatically

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

calculates the square root of the values before they are injected, such that in summary the temperature rise again scales linearly with the chosen voltage value. For simplicity, we will refer to the above written values as the injected voltage values, even though in reality their square root is injected into the heater. One volt that is injected to the heater does usually not create a temperature pulse in the thermometer that is measured with a height of 1 V by the SQUID. Therefore we normalized the measured voltage values for better readability, i.e. we divided all pulses by the height of the 1 V heater pulse. We then divided every pulse by its injected voltage value. To summarize, we performed on a test pulse $P_v[x]$ that originated from an injected volt value v into the heater the transformation:

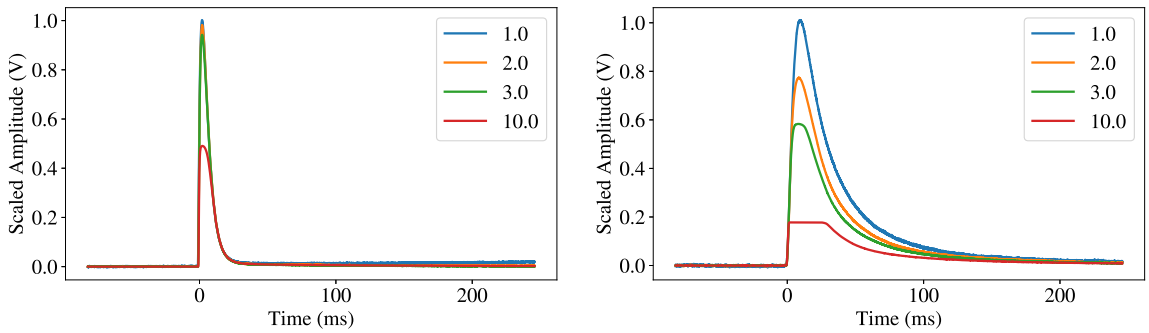
$$P_v[x] \leftarrow \frac{P_v[x]}{h_{1\,\mathrm{V}} \cdot v}, \tag{3.17}$$

where h_{1V} is the height of the test pulse that originated from the 1 V heater pulse. If our assumption holds, all pulses should overlap in the non-saturated regime, i.e. in the front part of their rising edge and in the rear part of their tail. This is the case for the saturated test pulses of the TUM40 phonon detector, as we can see in Fig. 3.9a. The 1 V (blue), 2 V (orange) and 3 V (green) test pulses overlap almost completely, except for their very peak. The 10 V (red) test pulse shows a stronger saturation but still overlaps with the other test pulses in the non-saturated regime. This proves our assumption that the pulses saturate only in height but not in width and renders the truncated fit an applicable method.

The situation is very different for the same scenario with test pulses from the Michael light detector, as we can see in Fig. 3.9b. Apart from a stronger saturation of all pulses, we also see a scaling in the width of all pulses. The 2 V pulse (orange) is more narrow than the 1 V pulse (blue), also in the non-saturated regimes. The same applies for the 3 V pulse (green) and even more for the 10 V pulse (red). Imagine the peak of the 10 V pulse to be reconstructed with a truncated fit – it would only reach a value of roughly 0.5 V. The energy reconstruction of pulses in this regime would therefore be biased of up to 50% of its true value.

This observation can only be explained by another non-linearity in our system, that originates not from the saturation of the TES transition curve. The most evident assumption we can make is that the thermometer loses heat by a temperature dependent process that we did not consider in our pulse shape model. This assumption seems plausible because the phonon detector operates in the bolometric mode, with an resistance change proportional to the heat that is transferred to its thermometers electrons, but the light detector in the calorimetric mode, where the resistance changes

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 3.9: Scaled 1 V (blue), 2 V (orange), 3 V (green) and 10 V (red) test pulses from (a) the TUM40 phonon detector and (b) the Michael light detector. The pulses were scaled with equation 3.17. (a) We can see an overlap of the pulses in the non-saturated regime. The pulses saturate only in their height, but scale linearly in their width. (b) We see no overlap between the pulses, they therefore scale also non-linearly in their width.

proportionally to the total deposited energy in the thermometer electrons. Therefore a loss of heat would in the bolometric mode only cause a saturation in the height of the pulse that mixes up with the saturation of the TES – we would not even notice the effect. In the calorimetric mode it would cause the observed narrowing of the overall pulse shape.

Self-heating, or *electrothermal feedback*, a process that we did not consider so far, does indeed introduce a temperature depended loss of heat in the TES. This is an effect that occurs in all electrical components that feature a temperature dependent electrical resistance, it was studied in Ref. [54] for CRESST detectors. The TES is operated in a parallel circuit with a shunt resistor R_{shunt} and the readout coil of the SQUID, that is in charge of the resistance measurements. The circuit diagram of this situation is shown in Fig. 3.10. As the changes in the TES resistance are the only time dependent component in the system and the coil is superconducting itself, we can ignore the coil for the calculation of the currents in this system. The system is powered with a constant bias current in the order of magnitude of $\mathcal{O}(\mu A)$, from which a part

$$I_{TES}(t) = \frac{I_{bias}}{1 + R_{TES}(t)/R_{shunt}} \tag{3.18}$$

runs through the TES, causing a power loss of

$$P_{TES}(t) = R_{TES}(t) I_{TES}^2(t) = \frac{R_{TES}(t) I_{bias}^2}{\left(1 + R_{TES}(t)/R_{shunt}\right)^2}. \tag{3.19}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

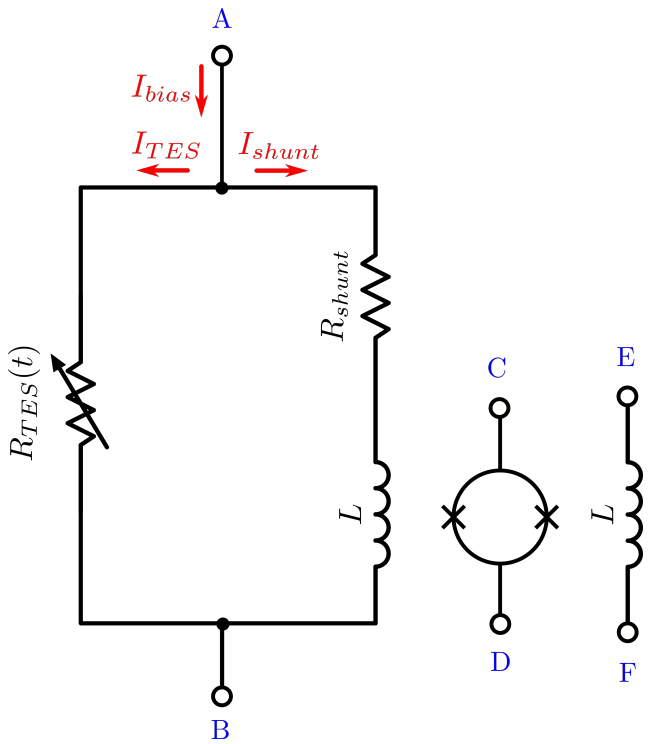


Figure 3.10: The circuit diagram of the TES. A constant bias current I_{bias} runs through the parallel circuit of the TES, the shunt resistor R_{shunt} and the readout coil L. The current running through the TES depends on its resistance and can be calculated with equation 3.18. Between the claps C and D the SQUID induces a voltage signal proportional to R_{TES}.

This power loss at the TES is at the same time a heat increase in the thermometer electrons and should therefore be considered in our derivation of the pulse shape model, i.e. we should adapt equation 3.5 accordingly to take the self-heating into account:

$$P_e(t) = \theta(t)(1-\epsilon)P_0 e^{-t/\tau_n} + P_{TES}(t). \tag{3.20}$$

As it happens, the TES of our detectors are typically operated in a temperature range where equation 3.19 is approximately linear with negative slope [54]. The self heating decreases inverse proportionally to the rise of the resistance of the TES, corresponding to a heat loss in the thermometer compared to the system in equilibrium. This is exactly the type of effect we were looking for to explain the narrowing of high energetic pulses.

All this said, if Michael's narrowing of high voltage test pulses is really and solely caused by the self-heating of the TES is still in question and a matter of ongoing research. It certainly needs further investigation before we can draw any meaningful

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

conclusions. Especially there is the need for a study of the test pulses from other bolometric and calorimetric detectors to investigate if they show the same effect.

Possible other factors, that were neglected by us, are the temperature dependence in the thermal resistances and heat capacities within the system and the current dependency of the superconductor. A current induces a magnetic field within the thin film and breaks down the superconducting state – a change in the TES bias current due to self-heating therefore causes a deviation in the form of the transition curve. This was already simulated for superconducting thin layers in atom chip experiments and might also contribute to the form of the measured pulse shape [55].

However, if the effect is indeed linked to the calorimetric regime then an improvement of the pulse shape model in this direction could have significant impact for the high energy reconstruction and the truncated fit in CRESST-III, where all detectors are calorimeters.

To conclude, there is at the moment no model to simulate the saturation for high energetic pulses reliably, as long as the question of Michaels narrow test pulses is not clarified. We therefore do not consider saturated events in the recoil event simulations of this work. As CRESST is interested in low energetic recoil events anyway, this does not impair our methods or results.

For the future it would certainly be interesting to improve the pulse shape model even further. A more detailed thermodynamic model might not yield an analytical template for the pulse shape anymore, however the coupled differential equations (3.2), (3.3) could still be solved numerically. Also, a simulation of the TES transition curve, depending on the temperature operation point and the TES bias current, could be included to describe the saturation. For the application of machine learning methods better capabilities to simulate data directly lead to better capabilities to analyze similar data automatically and reliably.

In summary, we introduced in this chapter a model, motivated by thermodynamics, to simulate the typical pulse shape of recoil events. We also introduced a method to simulate noise baselines with a given noise power spectrum and drift structure. The superposition of the simulated pulses and baselines provides us with realistically simulated recoil events that we can use to train machine learning models. In the following three chapters, we do exactly that. In chapter 4 we train a supervised regression model to predict the energy and onset timing of recoil events. In chapter 5 we apply different unsupervised models for the visualization of our events and the detection of anomalies. Finally in chapter 6 we train a supervised classification model

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

to distinguish between recoil pulses in the absorber and the carrier crystal, as well as between low energetic recoils and noise. For every method we compare our results with the knowledge-based method that are currently used in the analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Chapter 4

Pulse Height and Onset Timing Regression

In the previous chapter we studied the simulation of recoil events by deriving a model for the characteristic pulse shape of their voltage trace and a method for the simulation of noise baselines. One task of the raw data analysis is solving the inverse problem to the simulation: The estimation of the energy and the timing onset for the given raw data of a recoil event. This is a regression problem. The amplitude of the pulse scales linearly with the deposited energy from the recoil. The precise reconstruction of the pulse height is therefore necessary for a low trigger threshold and high quality of the dark matter limit.

The raw data event is a voltage trace, represented by a numerical array with 8192 elements. As a time series, the samples of the raw data event are strongly autocorrelated, i.e. the value of a sample strongly depends on the previous samples. Deep learning, i.e. machine learning with neural networks, is well suited to solve high dimensional and highly correlated regression problems. To keep our work accessible for non-experts for neural networks, we introduce the technical details of various neural networks in appendix A and stick in the main text to generalized and qualitative descriptions of the internal workings of our models.

Currently for the height regression task in CRESST analysis either the optimum filter or the standard event fit is used [39]. The standard event fit is used for the reconstruction of the pulse onset timing as well, which is a useful utility for discriminating tools. In the first section of this chapter we study the optimum filter and its performance for height regression on the raw data events. In the second section, we introduce the standard event fit and its results for height and onset timing regression. Is the last section, we introduce our deep learning model for the tasks of height and timing regression and compare the results of all three methods.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

4.1 The Optimum Filter

The voltage trace of a recoil event is shown in Fig. 2.4a for a TUM40 phonon event and in Fig. 2.4b for a Michael light event. The true height of the pulse shape is distorted by the Gaussian noise of the detector. In a naive evaluation of the pulse height, where we take the highest value of the array as an estimate for the true pulse height, we have a bias in the order of magnitude of $\mathcal{O}(\mathrm{mV})$ compared to the true height. This is certainly an unwanted property, especially for low energetic pulses where the signal-to-noise ratio can be in the order of magnitude of $\mathcal{O}(1)$.

The simplest way to counter this issue is to perform a *moving average* on the time series, i.e. we calculate a new time series $\tilde{e}[x]$ in which every element is the average of N, a natural number, elements of the recoil event's time series $e[x]$:

$$\tilde{e}[x] \leftarrow \frac{1}{N} \sum_{i=x-N/2}^{x+N/2} e[i]. \tag{4.1}$$

Here x is the array index, chosen that way to avoid confusion with the time t. Typically the value $N = 50$ is chosen for the moving average. This imposes boundary effects in the first and last $N/2$ samples of the new time series, which is usually fixed by padding the original time series with an appropriate number of zeros on both sides. The moving average is interpretable as a simple kernel-estimation technique, which are commonly used in statistics [56, 57].

The effect of the moving average is that stochastic, high frequent fluctuations cancel out and the bias of the height estimate is reduced. However, there is still a bias towards higher values left that weakens results of the experiment. We can think of the moving average as a low pass frequency filter, which it indeed is [58]. Optimized frequency filters are currently the state-of-the-art method to improve the signal-to-noise ratio and are calculated by convolving a time series with a transfer function [59]. The application of an optimum filter to CRESST data was studied in Ref. [40]. The convolution can be realized numerically more efficiently as a multiplication in Fourier space, according to the convolution theorem. The optimum filter is a matched filter, optimized for the detection of a specific signal shape $s(t)$ overlaid by Gaussian noise with the noise power spectrum $\mathcal{N}(\omega)$. The transfer function of the optimum filter in Fourier space is given by

$$H(\omega) = h \frac{s^*(\omega)}{N(\omega)} e^{-i\omega\tau}, \tag{4.2}$$

where $s^*(\omega)$ is the complex conjugate of the typical signal shape's Fourier transform $s(\omega)$, ω the coordinate in Fourier space and h a normalization constant, preserving the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

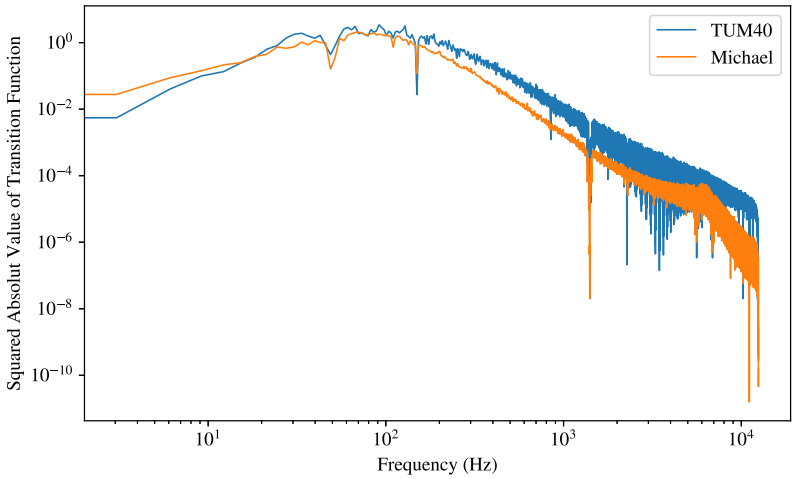


Figure 4.1: Power spectrum of the optimum filter transition function for TUM40 phonon events (blue) and Michael light events (orange). A reduction of a factor > 10 is visible for frequencies below 20 Hz and above 200 Hz, as well as for the mains voltage value 50 Hz and its second harmonic 150 Hz.

height of the signal shape. τ is a parameter to adjust the delay of the filter, typically chosen as the time of the peak of the signal shape. The normalization constant can be calculated with

$$h = 1 \Big/ \int \frac{|s(\omega)|^2}{\mathcal{N}(w)} d\omega. \tag{4.3}$$

These calculations can be performed numerically via the fast Fourier transform and an appropriate numerical integration technique [60, 61]. The optimum filter is indeed the optimal frequency filter for the detection of a specific signal, i.e. it maximizes the signal-to-noise ratio. A comprehensible derivation of this statement is given in Ref. [62]. However, the optimum filter does not preserve any information about the timing of the signal shape, i.e. the peak of the filtered time series can show a significant temporal deviation compared to the peak of the unfiltered signal.

For the height reconstruction of TUM40 recoil events, we use our fitted pulse shape model as signal shape $s(t)$ that we calculated in chapter 3 together with the noise power spectrum. The power spectrum of the calculated transition function is shown in Fig. 4.1, for recoil events from the TUM40 (blue) and the Michael (orange) detectors. We can see a reduction of frequency amplitudes of at least a factor 10 below 20 Hz and above 200 Hz, as well as at 50 Hz and 150 Hz. Latter ones are caused by the high amplitude of the frequency of the mains voltage and its second harmonic in the noise power spectrum. The dominant frequencies of the pulse shape around 100 Hz are not reduced. However, Fig. 4.1 shows only the reduction of frequencies while the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

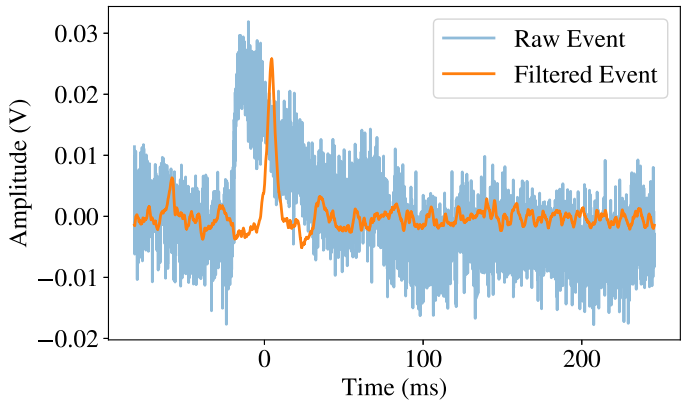


Figure 4.2: Simulated recoil event of the TUM40 phonon detector (blue) with true height 25 mV. The output of the optimal filter (orange) does not resemble the characteristic pulse shape, but reconstructs the true pulse height very well.

information about the local structure, the shape of the sought-for signal, is contained in the phase information, given by the imaginary part of the transition function.

A filtered TUM40 recoil event, that was simulated with the methods introduced in chapter 3, is shown in Fig. 4.2. In the raw data event (blue) the true pulse height of 25 mV is strongly distorted by the Gaussian detector noise. The filtered event (orange) does not resemble the characteristic pulse shape, but reconstructs the true pulse height very well. We can see that the peak of the filtered signal shows a temporal deviation from the peak of the raw event.

To test the performance of the optimum filter we simulate data sets of TUM40 phonon pulses with true heights of 150 mV, 110 mV, 80 mV, 60 mV, 45 mV, 25 mV, 6.8 mV, 3.4 mV, 1.2 mV and 0.6 mV. For each height a sample of 1000 pulses is simulated, resulting in a data set with 10000 samples in total. We apply the optimum filter to reconstruct the height of the pulses. The results are documented in Tab. 4.1. We see a reasonable match of the sample means down to a true height of 6.8 mV with a standard deviation of roughly 1 mV. For smaller pulses the reconstructed pulse heights are strongly biased, due to the detector noise.

The result of a χ^2-test supports the assumption of a Gaussian error distribution around the sample means for pulses of 25 mV and above, but rejects it for lower values. The test is done with 8 equidistantly spaced bins in the interval $[\mu - \sigma, \mu + \sigma]$ and one overflow and underflow bin. The last *ventile*, i.e. the value at which the cummulative distribution function equals 0.95, of the χ^2 distribution with seven degrees of freedom is 14.07, which is the value above which we have to reject the corresponding hypothesis when we extract two parameters from our data [63]. However, if we want to test the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Amplitude (mV)	μ (mV)	σ (mV)	χ^2-test with true amplitude	χ^2-test with μ
150.0	150.11	1.11	**17.74**	6.57
110.0	110.09	1.07	11.32	10.78
80.0	80.05	1.08	4.63	2.73
60.0	60.06	1.13	5.82	6.45
45.0	45.04	1.06	1.85	2.4
25.0	25.07	1.05	8.88	8.56
6.8	7.06	1.01	**57.62**	1.52
3.4	**4.11**	0.82	472.62	**44.99**
1.2	**3.37**	0.56	3214.33	**61.96**
0.6	**3.35**	0.57	3891.67	**58.65**

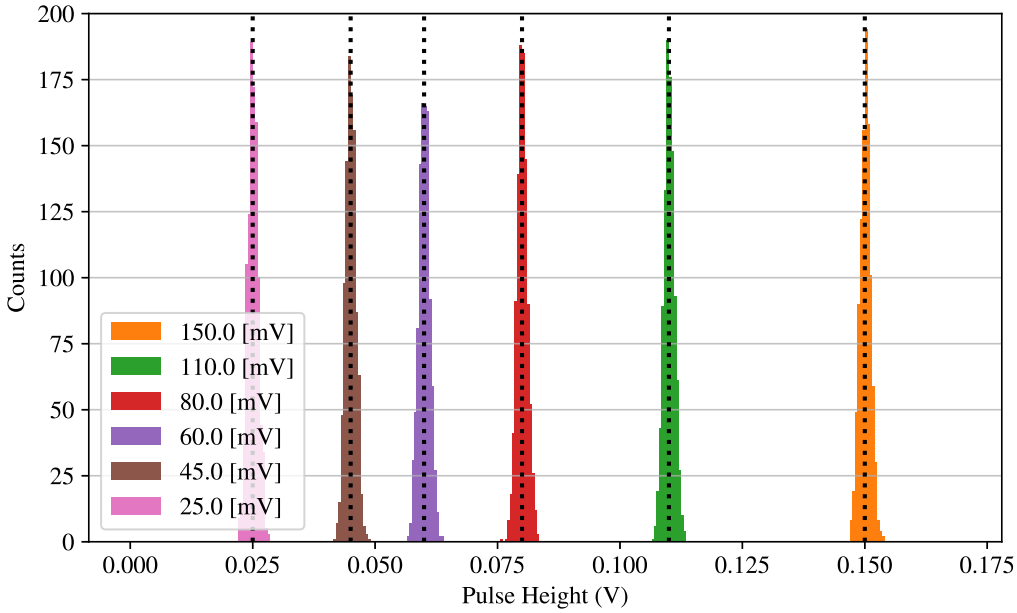
Table 4.1: Height reconstruction of simulated TUM40 phonon pulses with the optimum filter. For every amplitude a sample of 1000 pulses was simulated and the sample mean μ and sample standard deviation σ are determined. A χ^2-test rejects the hypothesis of a Gaussian error distribution around the sample mean for pulses of 6.8 mV and below. The value above which we have to reject the hypothesis is 14.07. The hypothesis of a Gaussian error distribution around the true amplitude has to be rejected for two more true amplitudes. The value above which we have to reject the hypothesis in this test is 15.5. For better readability of the table, values that indicate a mismatch between our hypothesis and the data are bold-faced and values of the chi^2-test with the true amplitude that were expected to be high, due to the other χ^2-test, are printed in gray.

hypothesis of a Gaussian error distribution around the true amplitude, we have to replace the sample mean μ in the χ^2 test with the true amplitude. Also, as we only extract one parameter from the data, we have to compare the results of the χ^2-test with the last ventile of the χ^2-distribution with eight degrees of freedom, which has the value 15.5. This hypothesis has to be rejected for the true amplitudes 150 mV and 6.8 mV as well, which is due to the high sensitivity to a bias in the sample mean.

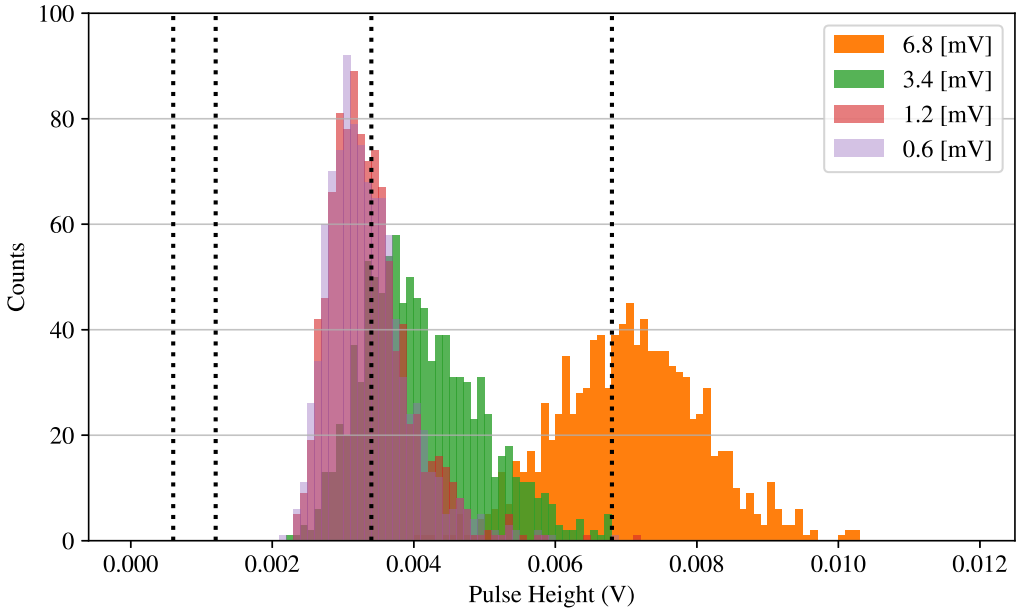
The trigger threshold for the analysis of the TUM40 phonon detector from run 33 was set to a recoil energy of 0.6 keV, which corresponds to roughly 5.4 mV. For our simulated pulses, this is also roughly the threshold below which the error distribution is not Gaussian anymore and the bias of the sample mean increases significantly – we can therefore confirm this threshold. For CRESST-III, where a continuous data acquisition is used, the optimum filter achieves usually significantly better trigger thresholds than the hardware trigger. However, in the run 33 analysis the optimum filter was not yet used, but instead the standard event fit that we will introduce in the following section.

The results from Tab. 4.1 for the pulses above 25 mV are visualized with histograms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a)



(b)

Figure 4.3: Histograms of the reconstructed height of simulated TUM40 phonon pulses (a) above 25 mV and (b) below 6.8 mV. For the reconstruction the optimum filter was used. The error distribution (in color) is approximately Gaussian around the true values (black, dotted) for values down to 0.6 mV. Below this value, the sample means are strongly biased and the distribution is not Gaussian anymore.

in Fig. 4.3a. We see the true heights of the simulated pulses as black dotted lines, surrounded by the histograms of the corresponding reconstructed pulse heights (in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

color). We can see the approximately Gaussian error distribution. In Fig. 4.3b the pulses with heights below 6.8 mV are visualized. The histogram shape of the pulses below 3.4 mV is strongly asymmetric and biased towards higher values.

To conclude, in this section we introduced the optimum filter, which is the optimal frequency filter to improve the signal-to-noise ratio of a characteristic signal shape. For our tests on simulated TUM40 pulses, we achieved a almost unbiased pulse height reconstruction, with a standard deviation of roughly 1 mV, down to a threshold of roughly 6.8 mV. This threshold will certainly vary for different detectors, due to a different pulse shape and noise background. The TUM40 detector, that was used in run 33, has an energy-voltage conversion (EPC) rate of 111.24 keV/V, therefore our voltage threshold corresponds to an energy threshold of 0.75 keV for the unbiased pulse height reconstruction. In run 33, the hardware trigger was set to a threshold of 0.6 keV, which seems reasonable, if we accept a slight bias for low energetic events. However, in run 33 the optimum filter was not yet used, but the standard event fit as a method for the pulse height reconstruction (see section 4.2). The optimum filter was first used in the analysis of detector A, used in CRESST-III run 34, and achieved a threshold of only 30.1 eV, which corresponds with detector A's significantly lower EPC rate of 4.38 keV/V to a voltage threshold of 6.87 mV. This match of detector A's voltage threshold with the value we could reconstruct is a coincidence. The test we did on TUM40 data is not directly compatible with other detectors, due to the individual pulse shapes and noise power spectra. Detector A has a lower EPC rate than TUM40 because all detectors in CRESST-III are roughly 10 times lighter than in CRESST-II. In the following section we introduce the standard event fit, that was used for the height reconstruction in run 33 and before. It is also used for the evaluation of the onset timing of pulses.

4.2 The Standard Event Fit

A typical pulse shape of a detector crystal's recoil event can be constructed by averaging several hundred pulses of similar height. In doing this, the stochastic noise cancels and a clean pulse shape is left. To this pulse shape, we can fit our pulse shape model, as we did in Fig. 3.2a. To reconstruct the height and onset timing of a raw data event, we can use the pulse shape model with most parameters fixed and fit its time shift, height and baseline offset and drift to the raw data event. The fitted height and timing onset parameters resemble the true values.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

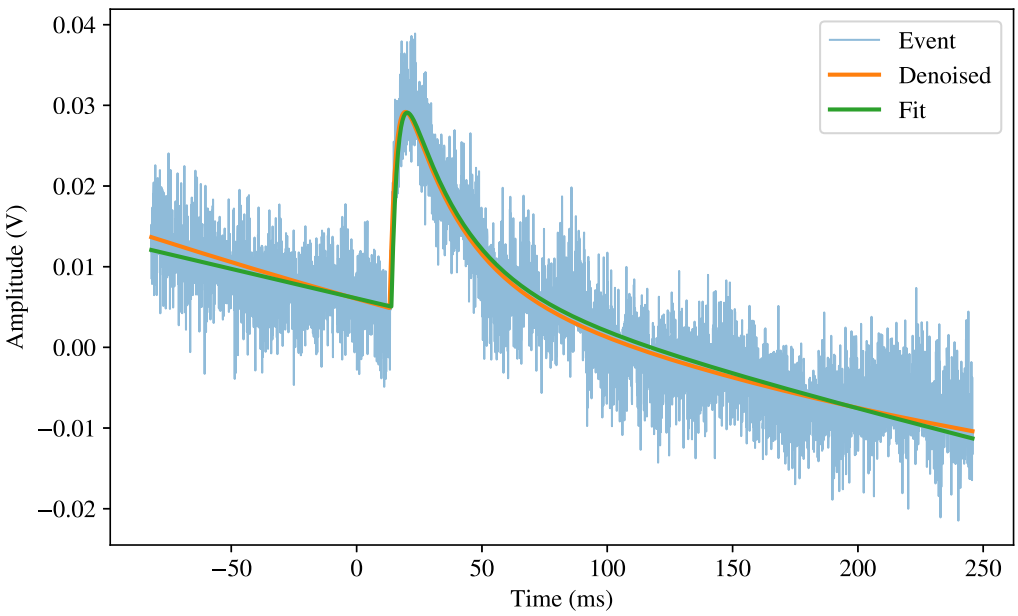


Figure 4.4: Standard Event Fit (green) on a simulated TUM40 phonon event (blue) and the denoised event, i.e. its pulse shape model together with the baseline drift (orange). We see a relatively good match of the fit and the denoised event, which indicates a good reconstruction of the true height and timing onset by the standard event fit.

As fitting model, we use a pulse shape superposed with a baseline offset and linear drift. We obtain the fit parameters by solving the minimization problem

$$\{\hat{h}, \hat{t_0}, \hat{k}, \hat{d}\} \leftarrow \underset{h,t_0,k,d}{\arg\min} \sum_x \left(e[x] - h \cdot s(t - t_0)[x] + kt[x] + d\right)^2, \tag{4.4}$$

where $t[x]$ is an array containing the time values corresponding to the array indices, $e[x]$ is the raw data event, $s(t - t_0)[x]$ is a function that returns the time series of a TUM40 pulse shape with onset at t_0. The fit parameters h, t_0, k and d describe the height and onset timing of the pulse shape and the linear drift and offset of the baseline.

We see an example of the standard event fit in Fig. 4.4, applied to a simulated TUM40 phonon event (blue). The fit (green) matches the denoised event, i.e. its pulse shape model together with its baseline drift, relatively well. This is an indication for a good estimate for the pulse height and the onset timing of the event.

To test the performance of the fit, we use the same data set as in the last section, a data set of simulated TUM40 phonon pulses with true heights h of 150 mV, 110 mV, 80 mV, 60 mV, 45 mV, 25 mV, 6.8 mV, 3.4 mV, 1.2 mV and 0.6 mV, with a sample size of 1000 pulses each. The timing onset t_0 of the simulated pulses is sampled from

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Amplitude (mV)	μ_h (mV)	σ_h (mV)	χ^2-test with true amplitude	χ^2-test with μ_h	μ_{t_0} (ms)	σ_{t_0} (ms)
150.0	149.93	1.77	5.49	6.15	-0.01	0.04
110.0	109.91	1.71	8.03	7.62	-0.01	0.05
80.0	79.97	1.71	5.65	7.3	0.0	0.07
60.0	59.93	1.68	5.04	8.26	0.04	0.09
45.0	44.89	1.73	**15.62**	9.34	0.03	0.17
25.0	24.77	2.09	42.97	**55.76**	0.13	0.37
6.8	**5.83**	**3.07**	202.29	**157.4**	**5.06**	**26.52**
3.4	**2.73**	**6.11**	1026.81	**986.03**	**12.71**	**65.31**
1.2	**2.03**	**5.96**	1197.28	**1145.85**	**22.53**	**71.18**
0.6	**1.83**	**9.15**	1954.7	**2001.56**	**25.04**	**81.85**

Table 4.2: Resulting sample mean and sample standard deviation of the fit parameter from a standard event fit with a sample size of 1000 for each amplitude. The values from χ^2-tests for the hypothesis of a Gaussian error distribution of the reconstructed pulse heights around the sample mean and the true amplitude reject the hypothesis for amplitudes smaller than 25 mV and 45 mV, respectively. The corresponding last ventiles of the χ^2 distribution are 14.07 and 15.5. The onset timing is for high amplitudes correctly reconstructed, but fails for amplitudes below 6.8 mV almost completely. For better readability of the table, values that indicate a mismatch between our hypothesis and the data, or a bad performance of the method, are bold-faced and values of the chi^2-test with the true amplitude that were expected to be high, due to the other χ^2-test, are printed in gray.

a uniform distribution in the interval $[-20\text{ ms}, 20\text{ ms}]$. The sample mean standard deviation, i.e. the standard deviation of a random number from the distribution divided by the square of the sample size, is 0.11 ms for our samples with size 1000. We therefore expect the true mean value of the onset timings in each sample in the interval $[-0.11\text{ ms}, 0.11\text{ ms}]$. The standard event fit is applied to all pulses, the reconstructed heights and timing onsets are documented in table 4.2. We see a good match of the sample means and reasonably low standard deviations for pulses with an amplitude above 25 mV. For the pulses with amplitudes of 6.8 mV and below, the sample means are significantly biased and the standard deviations increase to unreasonably high values. The high values of t_0's standard deviation implies that many standard events do not fit to the small pulses, but to some larger background fluctuation instead – otherwise a standard deviation higher than the interval of the uniform distribution, would not be possible.

As in the last section, we test the hypothesis of a Gaussian error distribution of the height reconstruction around the sample mean and the true amplitude, respectively, with a χ^2-test. The binning is again chosen as 8 equidistantly spaced bins in the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

interval $[\mu_h - \sigma_h, \mu_h + \sigma_h]$ and one underflow and overflow bin each. The resulting values of the χ^2-test have to be compared to the last ventiles of the χ^2 distribution with 7 and 8 degrees of freedom respectively, whose values are 14.07 and 15.5. The results support the hypothesis of a Gaussian error distribution around the sample mean for values above 45 mV and reject it for lower values. The assumption of a Gaussian error distribution around the true value is also narrowly rejected for a pulse height of 45 mV.

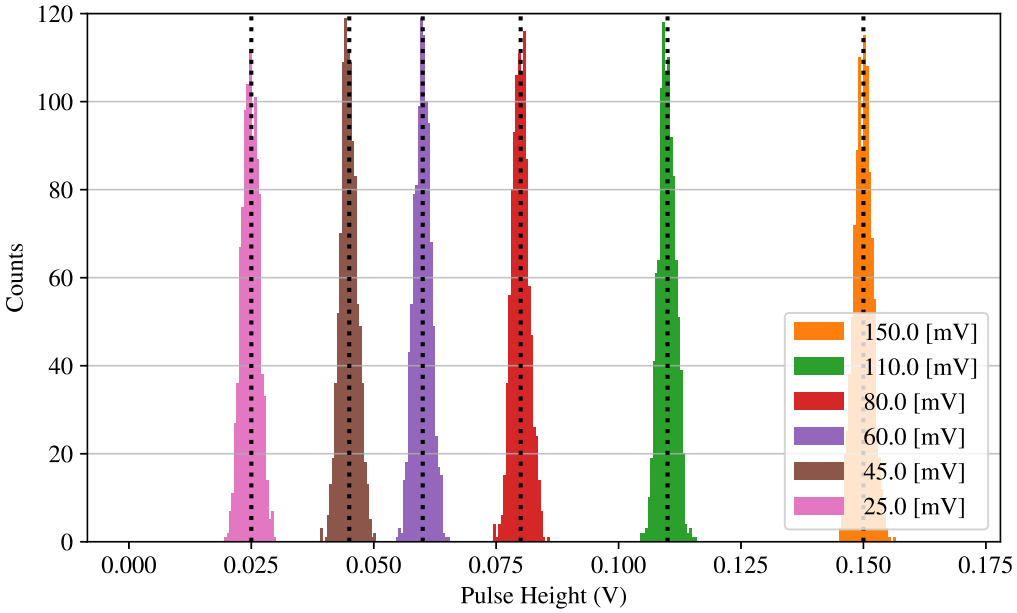
The results of the height reconstruction are visualized in Fig. 4.5a for pulse heights of 25 mV and above and in Fig. 4.5b for smaller pulses. The histograms of the high pulses are approximately Gaussian, for the smaller pulses no meaningful statement is possible. A comparison with the corresponding histograms from the height reconstruction with the optimum filter, see Fig. 4.3a and Fig. 4.3b, shows an overall higher standard deviation.

In the analysis of run 33, a threshold of 0.6 keV, corresponding to a pulse height of roughly 5.4 mV was achieved with the standard event fit [36]. In our test of the standard event fit on simulated TUM40 events, reconstructed pulse heights of 6.8 mV and below are significantly biased and feature a far higher standard deviation than larger pulses. Therefore, at this point, we cannot reproduce reasonably well the height reconstruction with the standard event fit to confirm the TUM40 threshold from run 33. However, Ref. [31] provides a detailed documentation of the techniques, that lead to this low threshold. The standard event fit can be improved with a more complex baseline model, namely with quadratic and cubic drift components or a quadratic and an exponentially decaying component with a decay time fixed to the decay time of control pulses. Also, we did not constrain the minimization problem that we solve in our fit. The results could most probably be improved with clever constraints on the parameters. A third difference from the data we analyzed to the data from run 33 is, that we tested only simulated data. It is possible, that the standard event fit might show statistically slightly different results on the originally measured data from the TUM 40 detector, used in run 33. Therefore it is reasonable to assume that the standard event fit, if appropriately optimized, might achieve a threshold of $\sim$ 5-6 mV.

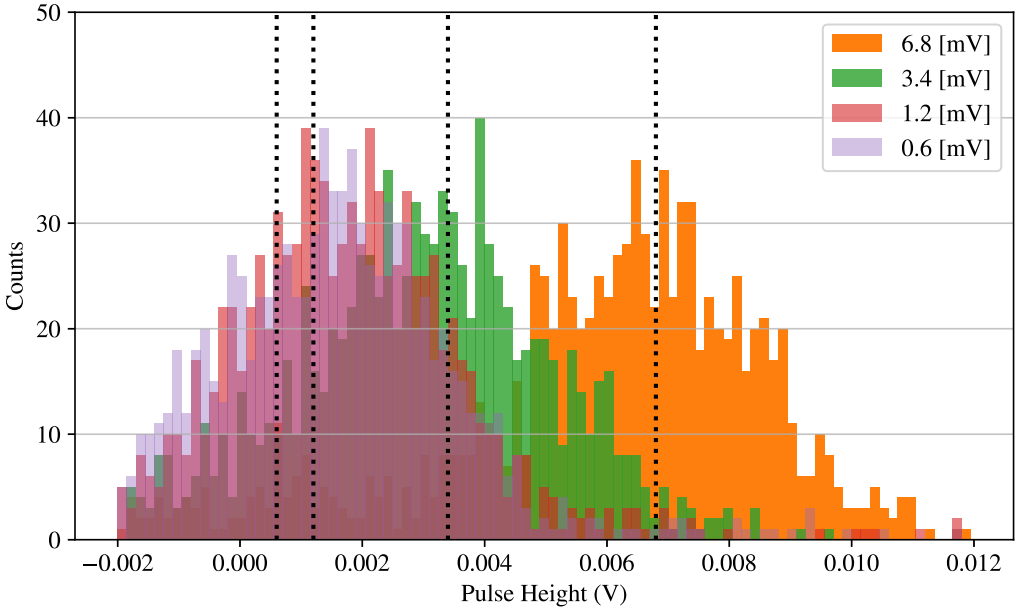
We can certainly confirm a precision in the onset timing detection of roughly $60\mu s$, as claimed in Ref. [31].

To conclude, we have reached significantly better results with the optimum filter than with the standard event fit for the task of height regression. For the pulse onset detection, the standard event fit shows very good results. The standard event fit can

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub



(a)



(b)

Figure 4.5: Histograms of the reconstructed height of simulated TUM40 phonon pulses (a) above 25 mV and (b) below 6.8 mV. For the evaluation of the height the standard event fit was used. The error distribution (in color) is approximately Gaussian around the true values (black, dotted) for values down to 25 mV. Below this value, the sample means are strongly biased and the distribution is not Gaussian anymore.

be used for the detection of anomalies as well, as we will study in chapter 5 in detail. In the following section we will introduce a deep learning model for the tasks of height

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and timing regression and compare its results to those of the optimum filter and the standard event fit.

4.3 Deep Regression

Deep learning, i.e. machine learning with *neural networks*, has gone through a phase of rapid development in the past ten years. The most famous examples are its application to image and voice recognition, which are today used by every smartphone. Deep learning is not only extremely efficient in classification problems, but also in regression problems [64]. It is increasingly used in particle physics and astrophysics, e.g. for the detection of gravitational waves [65], detector unfolding [66], the search for physics beyond the standard model [67, 68] and also the search for dark matter [12].

A neural network is a *function approximator*, i.e. for a given data set of numerical samples x_i and corresponding numerical labels y_i it can learn to approximate a function $\hat{f}$, that assigns a label to every sample:

$$\hat{f}(x_i) \approx y_i. \tag{4.5}$$

Neural networks are a special case of *supervised learning*, that we introduced at the beginning of chapter 3. For our pulse height and onset timing regression problem, the sample corresponds to the numerical time series of the raw event and the labels are the true height h and onset timing t_0. We only use simulated data in this work, therefore the labels are automatically given for every sample in our data set.

A thorough introduction in neural networks can be found in Ref. [69] and [70]. We introduce in this chapter only the basic concepts and vocabulary, and refer to appendix A for technical details in order to keep the main focus of this work on dark matter physics, not data science.

The most simple version of a neural network for regression $\mathcal{R}$ consists of two weight matrices L_0, L_1 and two bias vectors b_0, b_1 of appropriate shape together with an *activation function* σ. It maps the values of a sample to a prediction of its label:

$$\mathcal{R}(x_i) := L_1 \left(\sigma \left(L_0 \left(x_i\right) + b_0\right)\right) + b_1 \approx y_i. \tag{4.6}$$

The activation function is a function that introduces a non-linearity to the neural network while preserving most information contained in its input. Usually this is a piecewise smooth function with Lipschitz constant ≤ 1. A prominent example, that we also mainly use in this work, is the *rectified linear unit (ReLU)*, that equals the identity function for values ≥ 0 and zero for values < 0. The values in the weight matrices and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

bias vectors are called *weights* and are fit coefficients, solely determining the output of the neural network. Compared to e.g. a function approximation with polynomials, the neural networks typically has by far more degrees of freedom. However, the main advantage over most other approximation methods is the fast run time of neural networks. All calculations can be done for many samples in parallel. This provides the practical possibility to approximate more complex functions accurately. A sketch of the architecture of a neural network can be found in Fig. A.1.

There are many ways to improve the architecture of the simple neural network $\mathcal{R}$ that we introduced above. One way is to add more weight matrices and bias vectors with their corresponding activation functions. The pair L_i, b_i is called the i'th *layer* of the network. Modern architectures often have tens or even hundreds of layers. The intermediate state $l_i = L_i(\cdot) + b_i$ between layers is called the *hidden state* or *latent state*. It is also common to optimize the length of the vectors, that represent the hidden states, i.e. the number of *nodes*.

The quality of the weight optimization to a given approximation problem fully determines the performance of the network. These weights are optimized in a process called *training* or *learning*, which is an iterative algorithm: The network is initialized with randomly chosen weights and predicts values for all samples in the data set. The error between the predictions and the true labels is measured by a *loss function*, for which we typically use the mean squared error function. This error is then minimized with the *gradient descent* algorithm or a variation of it, e.g. the ADAM algorithm [71]. The time efficient optimization of all weight layers is accomplished by an algorithm called *backpropagation* [72]. The *learning rate* is a parameter between 0 and 1 and controls the magnitude of a weight update. We call the process of prediction, loss evaluation and optimization for the whole data set an *epoch*, the training process typically consists of several tens of epochs, up to several hundreds and more. Usually the weight update is not calculated for the whole data set at once. Instead the data set is split into *mini batches*, consisting of several dozens of samples, and the weight update is performed for one mini batch at a time. For our work we chose a mini batch size of 64 samples.

Parameters like the learning rate, the batch size or the number of layers and nodes are called *hyperparameters*. A lot of research is invested in the optimization of hyperparameters [73], which is a laborious activity. Training a neural network usually needs a time frame between several hours to several days. Repeating this process for a different set of hyperparameters is time and cost intensive. For many applications,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

hyperparameters are optimized by hand following an educated gut-feeling rather than with an automatized algorithm.

Overfitting is defined as weak generalization from the training samples to similar samples. To avoid it, the data set is split in three parts: The *training set*, the *validation set* and the *test set*. While the training set is used to optimize the weights, the loss value of the predictions for the validation set is evaluated after every epoch to provide a reference for the performance of the network on yet unseen samples. The performance of different neural network models is usually compared by the error on the validation set. However, final metrics of the performance, i.e. statistics for representative types of samples, are for the final model calculated on the test set. While the train and validation set are typically split randomly, the test set is never merged with the others and left untouched to provide an unbiased performance testing. The composition of the test set is usually done such that meaningful statements about the performance of the network in a realistic setting are possible.

We want to train a neural network to predict the pulse height and onset timing of the TUM40 phonon events. We are especially interested in low energetic recoils, i.e. small pulses. We therefore simulate a data set, consisting of 27000 pulses with heights uniformly distributed in the interval $[0\ \mathrm{mV}, 150\ \mathrm{mV}]$ and onset timings uniformly distributed in the interval $[-20\ \mathrm{ms}, 20\ \mathrm{ms}]$. We randomly shuffle the data set and then split it into a training set of 24000 samples and a validation set of 600 samples. As test set we use almost the same data set from the previous two sections, with an increased sample size: A sample size of 3000 of pulses with height 150 mV, 110 mV, 80 mV, 60 mV, 45 mV, 25 mV, 6.8 mV, 3.4 mV, 1.2 mV and 0.6 mV, i.e. altogether 30000 samples. The onset timings are distributed the same way as they are in the train and validation set.

Modern machine learning applications are mostly programmed in Python, with optimized and convenient-to-use libraries. We did so as well, by using the PyTorch library for our networks. Computing was done on two 20-core Intel Xeon Silver 4114 CPUs with 2.20GHz and four GeForce GTX 1080 GPUs with 12GB memory each.

Now that we introduced the basics about neural networks, the training process and the data we use, we now show the performance of different neural network models first for the pulse height regression problem, then for the onset timing regression problem.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

4.3.1 Pulse Height Regression

We introduced our training, validation and test sets and a simple neural network architecture in the beginning of this chapter. In this section we present the results from three different architectures of neural networks which we trained to predict the pulse height. They all have an improved architecture compared to the simple version from equation (4.6). The technical details of the architectures are explained in appendix A, we introduce them only qualitatively in this section.

The first model we used is a *convolutional neural network* (CNN). It convolves the sample with several weight kernels in parallel, before weights matrices are applied to the filtered sample. The kernels are optimized in the training process and we call a set of parallel kernels a convolutional layer. This type of network typically performs very well for spatial data, i.e. data with strong correlations between neighboring elements.

The second model is a *Long Short-Term Memory* (LSTM) network. It splits the sample in a sequence of lower dimensional vectors, which are recursively inserted into the network – one vector of the sequence is called a time step. The network has an internal memory and can store information for several hundred time steps. This architecture is especially effective for time-distributed data, i.e. for time series.

The third model is the *ResNet-18*, a small version of the ResNet, which is the current state-of-the-art architecture for image recognition. We made only slight adaptions to the model to apply it to a one dimensional (time series) regression problem, instead of a two dimensional (image) classification problem. It has a high number of convolutional layers with *skip connections* in between. This provides the network the option to optimize the number of layers it uses for a given problem.

For the CNN and LSTM model hyperparameters, especially the number of layers and nodes, were optimized on a trial and error basis. The results and the hyperparameters of the best performing architectures after 15 training epochs are shown in table 4.3. The training took roughly two hours for each model, training longer than 15 epochs did improve the results only slightly, but comes with a high computational cost. Longer training was therefore only performed for the final model, which we will study shortly. After every training epoch the validation loss value and model was saved. We can see that the LSTM had by far the best validation loss value over the 15 epochs. The CNN and LSTM performed better with only a few layers, which is probably due to the relatively simple structure of our data. The ResNet, which consists of many layers, performed significantly worse than the other two models.

In Fig. 4.6 the validation loss (blue dashed) and train loss (orange) values from the training of the LSTM model are visualized. The validation loss was evaluated after

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Model	Specifications	Learning Rate	Best Epoch	Best Validation Loss
CNN	2 Convolutional Layers, Kernel Size 8	0.001	14/15	0.00272
ResNet18	-	0.0001	14/15	0.05975
LSTM	3 Layers, Input Size 8, Hidden Size 80	0.0001	13/15	**0.000188**

Table 4.3: Results of 15 epochs training of three different neural network architectures for the task of pulse height regression. The validation loss value is the common metric to compare different models for the same task. The LSTM outperforms the ResNet and the CNN significantly, with a validation loss value more than an order of magnitude lower than those of its competitors. For the final performance testing on the test set, the training of the LSTM model was continued for 35 more epochs, reaching a final validation loss of 0.00005. The loss values are calculated with the mean square error loss function and are not physically interpretable, but solely provide a measure for the comparison of different network architectures.

every epoch, the training loss for every mini batch. We see that the loss value decreases exponentially, which indicates a well functioning training process. The fluctuations of the loss values between mini batches and epochs are typical for the ADAM optimizer. From the asymptotic extrapolation of the learning curve we can guess that the loss would not decrease much further for longer training. The validation loss is typically higher than the training loss, because the neural net was optimized on the training data. In our case they are roughly as high for every epoch, which indicates that our training and validation set are very similar and that our model generalized properly from the training data to similar data. Both aspects are a good sign and do not come as a surprise as the structure of our data is relatively simple.

For performance measurements on the test set, we continued the training of the LSTM for 35 more epochs. The validation loss decreased further by a factor of four, reaching a final value of 0.00005. The sample means and sample standard deviations of the LSTM's predictions on the test set are shown in table 4.4. We omitted the results of a χ^2-test in the table because the error distribution is clearly not Gaussian, which we will analyze shortly. We have to be very careful when we draw conclusions from the performance of a neural network. They are black box models and it is not always transparent why a model works well on an individual samples but does not on others. However, we see that the model is slightly biased in its predictions for all pulse heights, the bias increases for higher pulses. This effect is even stronger for the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

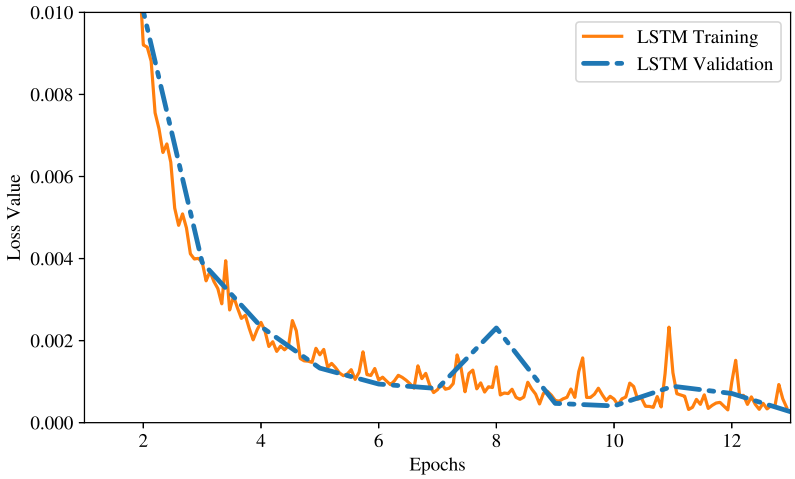


Figure 4.6: Learning Curves of the LSTM model for the training set (orange) and the validation set (blue). During the training, the loss of each mini batch is saved, while the validation loss is only saved for each epoch. Both curves have the characteristic form of an exponential decay and follow roughly the same trajectory, which indicates good generalization from the training to the validation set.

sample standard deviations, these increase by a factor of 60 from the 0.6 mV pulses sample to the 150 mV pulses sample. The better performance for smaller pulses is counter-intuitive and we will discuss this effect in the following paragraph. At this point there is one more eye-catcher to mention: The network performs incredibly well for small pulses. For pulses below 6.8 mV, the network has an roughly 20 times lower sample standard deviation than the optimum filter has for high pulses. It therefore seriously outperforms these knowledge-based methods for pulse height reconstruction.

The LSTM predictions on the test set are visualized in Fig. 4.7a for pulses higher than 25 mV and in Fig. 4.7b for smaller pulses. The predictions are shown as histograms (in color) around the true values (black, dashed). In both figures we can see a high precision of the prediction, indicated by the high histogram bin at the true value. For the small pulses, the predictions are roughly symmetrically distributed around the sample mean. The predictions for high pulses are not symmetrically distributed around the sample mean but feature a long tail towards lower values. This is indeed a strange behavior and not yet completely understood.

We believe that the counter-intuitively high error for large pulses is due to the non-linear numerical calculation of the neural network. All values in our time series range in the interval $[-10\text{ mV}, 10\text{ mV}]$, only the values of the peaks of high pulses range in a different order of magnitude and might therefore collide in some layer with non-linearities of activation functions that determine the predictions quality. Therefore

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Amplitude (mV)	μ (mV)	σ (mV)
150.0	**148.82**	**2.53**
110.0	**109.17**	**2.02**
80.0	79.5	**1.32**
60.0	59.66	0.89
45.0	44.78	0.59
25.0	24.89	0.29
6.8	6.8	0.08
3.4	3.42	0.06
1.2	1.24	0.04
0.6	0.64	0.04

Table 4.4: The sample means and standard deviations of the LSTM predictions for the test set. We can see a very good pulse height reconstruction for small pulses, but a bias of the sample mean and increasing standard deviations for high pulses. This is probably due to the numerical non-linearities in the architecture of the neural network, that weaken the models generalization for values of different orders of magnitude. For better readability of the table, values that indicate a high bias or standard deviation are bold-faced.

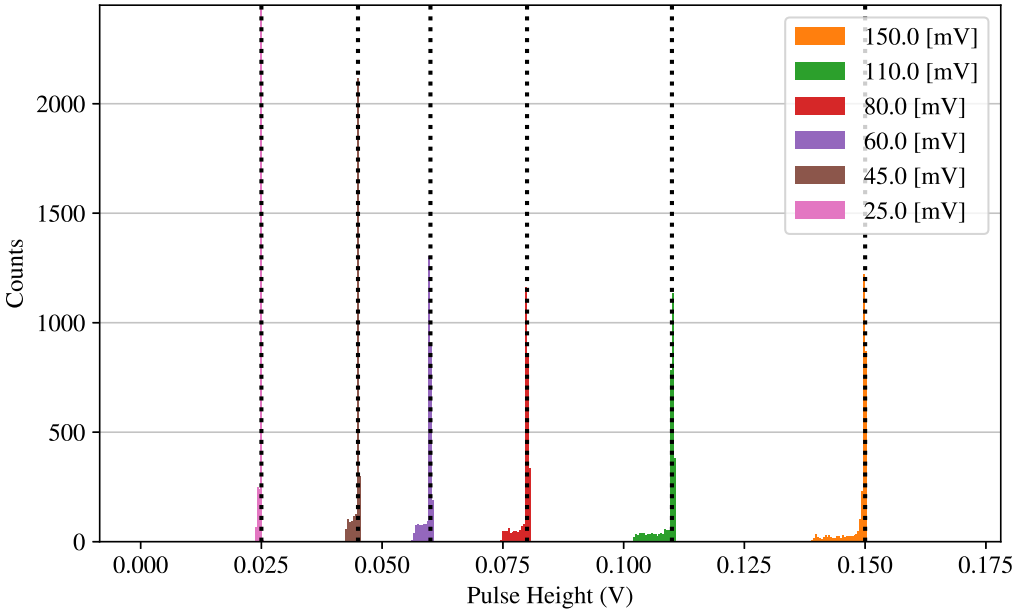
we believe that this problem could be improved by a clever normalization algorithm for the raw data samples or the individual training of an LSTM for every interval of values. This latter suggestion is called *ensemble learning* and often applied to improve the robustness of predictions [74].

In summary, we can reconstruct the pulse high of the TUM40 detector with an LSTM model down to 0.6 mV, which corresponds to roughly 65 eV. This is almost 10 times smaller than the trigger threshold of the TUM40 phonon detector in the run 33 analysis and smaller than most of the trigger thresholds of CRESST-III detectors, which have typically 10 times lighter crystals than TUM40. This result is indeed remarkable. However, there are still problems with the generalization to higher pulse heights. This could either be fixed by improvements to the neural network or by a clever division of labour between the LSTM and the optimum filter, that works very well for high pulses. In the following, we proceed with the application of the LSTM to the prediction of onset timings of pulses.

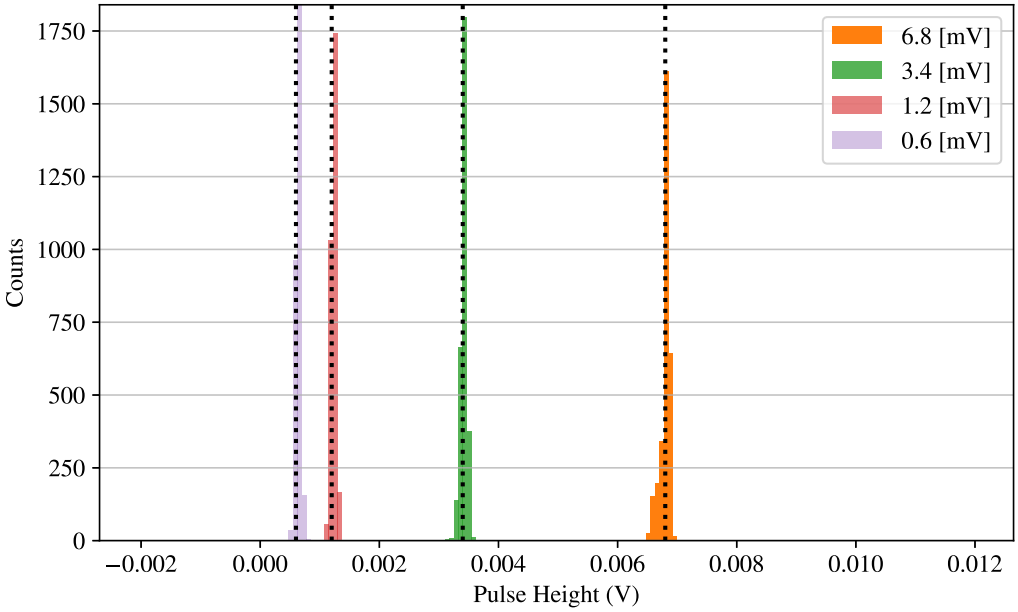
4.3.2 Onset Timing Regression

In this section we introduced the basics about neural networks and applied them to simulated TUM40 phonon pulses, yielding alluring results for determining their height, i.e. the recoil energy. Now we use them to determine the onset timing of the pulses. For

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a)



(b)

Figure 4.7: Histograms of the predicted height of simulated TUM40 phonon pulses (a) above 25 mV and (b) below 6.8 mV. For the evaluation of the height the LSTM model was used. The error distribution (in color) is roughly symmetrically around the true values (black, dotted) for values below 25 mV and features a small standard deviation. Above this value, the distributions around the sample means are asymmetric and show a long tail towards low values.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

this we use the model that performed best for the pulse height regression, the LSTM neural network. We also use the same simulated TUM40 phonon training, validation and test set as for the pulse height regression, i.e. pulse heights uniformly distributed between 0 mV and 150 mV and onset timings uniformly distributed between −20 mV and 20 mV.

Amplitude (mV)	μ_{t_0} (ms)	σ_{t_0} (ms)
150.0	-0.41	3.14
110.0	-0.65	2.36
80.0	-1.15	3.3
60.0	-0.36	3.19
45.0	-0.84	3.26
25.0	-2.82	4.18
6.8	-2.17	**11.06**
3.4	-2.34	**10.97**
1.2	-3.2	**12.36**
0.6	-2.11	**12.29**

Table 4.5: Sample means and standard deviations for the task of onset timing regression with an LSTM model. The true mean value for all pulse heights is (0 ± 1) ms. The standard deviation increases significantly for small pulses, indicating that the models predictions for the onset timing of small pulses does not hold any valuable information. For better readability of the table, values that indicate a bad performance of the network are bold-faced.

We trained the model for 15 epochs with a learning rate of 0.0001. The sample mean values and sample standard deviations of the LSTM predictions for the onset timings on the test set are documented in table 4.5. The true mean value of the onset timings in each sample is approximately (0 ± 0.1) ms. The predictions of the LSTM model are all biased towards low values. The bias and the standard deviation are about two orders of magnitude higher than the results from the standard event fit (see table 4.2). For small pulses, the standard deviation is roughly as high as the true standard deviation of the pulse onset distribution. The LSTM predictions are therefore nothing more than random numbers in the underlying interval.

To conclude, the standard event fit performs significantly better than the LSTM predictions for the onset timing regression problem. Maybe another neural network model would perform better for this task. However, deep learning does not perform better than knowledge-based methods for every problem, therefore we recommend the application of the standard events fit for the CRESST analysis, for now.

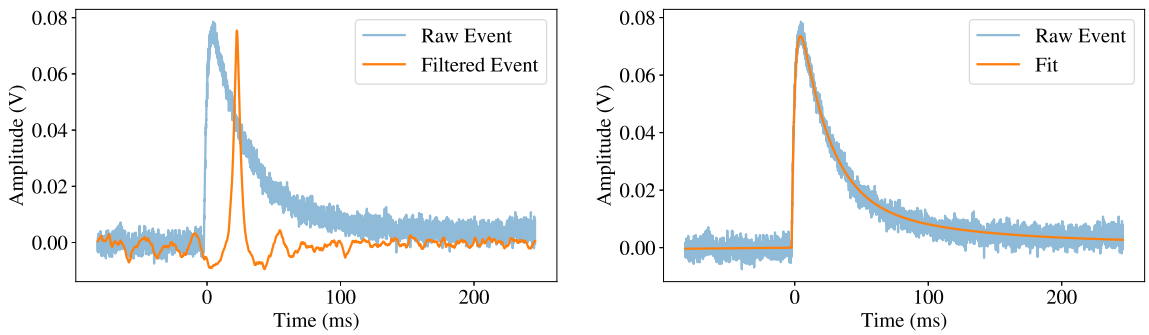
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

To summarize this section, we found a neural network architecture that shows extremely promising results on the pulse height regression of small pulses, with some issues for the prediction of high pulses that yet have to be fixed. For the task of onset timing regression, the standard event fit performs significantly better than the LSTM neural network. In this work we only used simulated data. In the last section of this chapter we evaluate the performance of the LSTM model that was trained on simulated data, in comparison to the optimum filter and the standard event fit, for the height reconstruction of some exemplary measured events from the TUM40 detector.

4.4 Results on Real Data

In the previous section we trained a neural network model on simulated TUM40 phonon events for the task of pulse height and onset timing regression. In this section, we present some exemplary results of the height reconstruction on measured raw data events from the TUM40 phonon detector. We compare the results from the optimum filter, the standard event fit and the neural network prediction for each event.

As a first example, we chose a regular recoil in the absorber crystal. It is shown in Fig. 4.8 (blue) together with the filtered event with the optimum filter (Fig. 4.8a, orange) and the standard event fit (Fig. 4.8b). The reconstructed heights of the events are 75.4 mV (optimum filter), 74.0 mV (standard event fit) and 76.6 mV (LSTM). They all are within a range of 1.6 mV, which is a reasonable deviation.



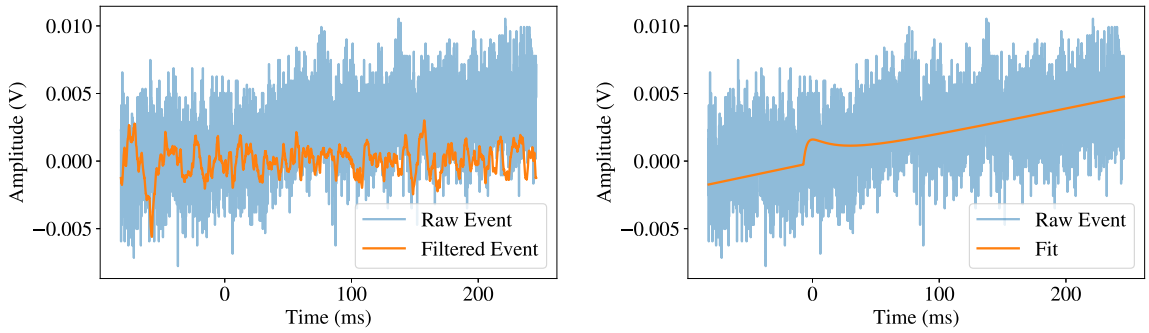
(a) (b)

Figure 4.8: Measured TUM40 phonon event (blue) (a) filtered with the optimal filter (orange), yielding a reconstructed height of 75.4 mV, (b) fitted with an standard event (orange), yielding a height of 74.0 mV. The LSTM model, trained on simulated events, predicts a height of 76.6 mV.

As a second example we chose an empty noise baseline. It is shown in Fig. 4.9 (blue) together with the filtered event with the optimum filter (Fig. 4.9a, orange) and the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

standard event fit (Fig. 4.9b). The reconstructed heights of the events are 3.0 mV (optimum filter), 1.7 mV (standard event fit) and −0.3 mV (LSTM). The reconstructed heights are all reasonably low, making it possible to define a trigger threshold below which events are classified as noise.



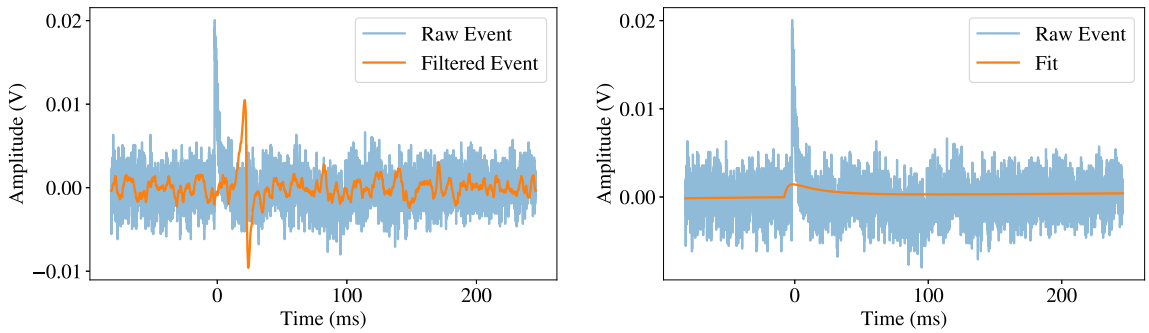
(a) (b)

Figure 4.9: Measured TUM40 phonon noise baseline (blue) (a) filtered with the optimal filter (orange), yielding a reconstructed height of 3.0 mV, (b) fitted with an standard event (orange), yielding a height of 1.7 mV. The LSTM model, trained on simulated events, predicts a height of −0.3 mV.

As a third example we chose a carrier event, i.e. a recoil in the carrier crystal. It is shown in Fig. 4.10 (blue) together with the filtered event with the optimum filter (Fig. 4.10a, orange) and the standard event fit (Fig. 4.10b). The reconstructed heights of the events are 10.5 mV (optimum filter), 1.5 mV (standard event fit) and 2.4 mV (LSTM). The reconstructed height values diverge strongly, which is a good indicator that the type of the event does not match with the standard event, the transfer function of the optimum filter and the simulated training set.
For the exemplary measured raw data events, the outputs of all three methods are reasonable. Either their values were close together, as for the recoil event and the noise baseline, or they predict diverging values, indicating that the event does not represent a recoil event in the absorber crystal and should be discarded. However, before our LSTM model is used to analyze the measured raw data of a CRESST run, a decent statistical study on a large set of measured raw data event should be conducted to test the networks generalization from simulated to measured data on a large scale.

To summarize, we introduced and evaluated the optimum filter, the standard event fit and an LSTM network for the tasks of pulse height and onset timing regression. For onset timing regression, the standard event fit shows the best results with a standard deviation below 100μs for pulse heights above 25 mV. The LSTM model predicts

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 4.10: Measured TUM40 phonon carrier event (blue) (a) filtered with the optimal filter (orange), yielding a reconstructed height of 10.5 mV, (b) fitted with an standard event (orange), yielding a height of 1.5 mV. The LSTM model, trained on simulated events, predicts a height of 2.4 mV.

pulse heights below 10 mV with a standard deviation of less than 0.01 mV, at least down to values of 0.6 mV. This is a significantly lower threshold than the optimum filter provides, reaching a typical CRESST-III trigger threshold with a CRESST-II detector. However, for higher pulses the LSTM model does show worse results. The optimum filter provides a robust height reconstruction with a standard deviation of roughly 1 mV down to a threshold of 6 mV.

Future work can build upon the results of this chapter by improving the LSTM architecture with a clever normalization algorithm or ensemble learning of different LSTMs for different pulse heights. Also, the results of the optimum filter and the LSTM could be combined to profit from the qualities of both. We applied our model to raw measured data in the last section of this chapter and also analyzed a typical non-absorber recoil event. In the following chapter we will study the raw, measured data in more detail and introduce several unsupervised learning methods for data visualization and anomaly detection.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Chapter 5

Unsupervised Techniques: Dimensionality Reduction, Anomaly Detection

We studied the simulation and energy reconstruction of recoil events in the previous chapters. The measured raw data of a CRESST run do not only consist of recoil events but of many artifacts as well, as accidentally recorded spikes from the electronics of the detector support system, SQUID resets and many more. We need to filter these events in order to calculate a high quality dark matter limit. In the current analysis chain this filtering is done by so-called *quality cuts*. These are discrimination methods to either discard a specific and well-known artifact or to discard every event that does not resemble the typical pulse shape of a recoil event. To provide a linguistic distinction between these types of cuts, we call the former one *specific* quality cut and the latter one *general* quality cut in the following. A good example for a general quality cut is the RMS-cut, that we study in section 5.3.

In this chapter we investigate unsupervised machine learning methods to improve the analysis and simulation. We apply two methods for dimensionality reduction to our events to visualize the raw data distribution. For this we introduce the *principal component analysis (PCA)* and the *t-distributed stochastic neighbor embedding (t-SNE)* and compare their results [75, 76]. In the section thereafter we use the standard event fit and *autoencoders*, a special neural network architecture, to detect anomalies in the data, i.e. events that do not resemble the characteristic pulse shape of recoil events. All four methods could be used to define general quality cuts, as they force events to cluster according to the similarities in their shape. The methods of dimensionality reduction can also be applied to explore the raw data distribution and possibly recognize event types that were yet unknown. It can therefore help to improve the simulation of raw

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

data by comparing the simulated data distribution to the raw data distribution.

We use a sample of 7000 consecutively measured events from the TUM40 phonon detector, that was used in run 33, as raw data. This corresponds to a measured time of roughly eighteen hours or an exposure of 0.2 kg days. The data acquisition automatically labels the events as baseline, test pulses or events, as we introduced it in section 2.3. However, roughly 2% of all baselines do also contain artifacts as spikes or squid jumps. Also, the recorded events are composed of recoil events and artifacts, the event label alone is therefore not satisfying to judge over the shape of their voltage trace. To provide a data set with reliable and descriptive labels, we introduce in the first section of this chapter a simulation of carrier events, recoils that happen inside the TUM40 carrier crystal. We then create a data set with a sample of 5000 simulated regular recoil events and carrier events each, with recoil energies between 0 mV and 150 mV. By this, we can test the discriminating power of our unsupervised methods on a labeled data set.

5.1 The Carrier Event

We introduced the design of CRESST-II detectors in chapter 2.2: A TES is mounted on a crystal, working as a highly sensitive thermometer (see Fig. 2.2). Every detector type has some unique design individualities. For the TUM40 detector, this is a separate, smaller *carrier crystal*, that is mounted in between the detector crystal and the TES. This design protects the larger crystal from high temperatures that are necessary for the evaporation of the TES. The two crystals are then glued together, which does not affect their internal structure. To avoid confusion, we call the larger detector crystal the *absorber crystal*. This design type, with separate absorber and carrier crystals, is called the *composite design*.

We can see a specific type of artifact in the raw data analysis, caused by the composite design: Recoil events that happen inside the carrier instead the absorber crystal. This event type can also be described with a pulse shape, but with a much faster rise and decay time than the absorber recoil events. Even though carrier events do origin from particle recoils, we cannot reliably calibrate their recoil energy and therefore have to discard them.

For the simulation of carrier events we can use the same procedure as introduced in chapter 2 for absorber recoils. Our derivation of the pulse shape model from section 3.1 is still valid. The only difference lies in the parameter ϵ, which describes the share of non-thermal phonons that decay in the TES. This parameter is much higher for a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

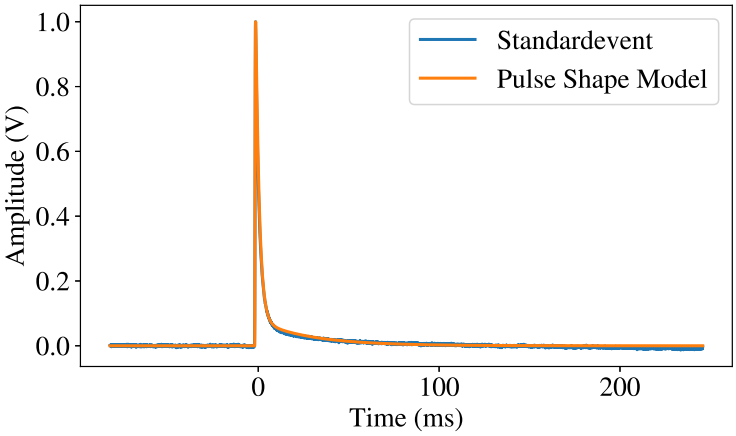


Figure 5.1: Standard event of a recoil in the carrier crystal (blue) and fitted pulse shape model (orange). The good match of their voltage traces indicate the capability of the pulse shape model to describe carrier events reliably.

carrier recoil than for an absorber recoil, resulting in a stronger first component of the overall pulse shape, i.e. a fast rise and decay time, combined with a long tail. We use the pulse shape model only as a fit function, therefore these differences do not affect our simulation technique.

First we construct a sample of several hundred carrier events of roughly similar height and superpose them. By doing so we construct the typical shape of a carrier event with high signal-to-noise ratio, i.e. a carrier standard event. Then the pulse shape model is fitted to the carrier standard event, by a minimization of the mean squared error with respect to the parameters of the model. The fitted pulse shape model and the standard event are shown in Fig. 5.1. The good match of the carrier standard event (blue) and the fitted pulse shape model (orange) indicates that the pulse shape model is indeed applicable for the description of carrier events.

For the final simulation of realistic carrier events we superpose their pulse shape with noise baselines, created with the method from section 3.2. We create a data set, consisting of 5000 carrier and absorber events each. The height of the events is uniformly distributed in the interval $[0\ \mathrm{mV}, 150\ \mathrm{mV}]$, the onset timing in the interval $[-20\ \mathrm{ms}, 20\ \mathrm{ms}]$. With this simulation we obtained a labeled data set that we can use to test the performance of the unsupervised methods, which we are going to introduce in the following two sections.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

5.2 Dimensionality Reduction

The numerical time series of an event consists of 8192 elements. We can therefore interpret an event as a vector in a 8192-dimensional vector space. The set of all elements in this high dimensional vector space is a sample from the underlying data distribution. To analyze this data distribution we try to map it to a distribution over a lower dimensional and better comprehensible vector space, while we try to preserve its characteristic properties. One way to do so is the light yield plot, that maps the energy information of the phonon and light channel of a detector in a two dimensional space. There are many more techniques to reduce the dimensionality of our data distribution, out of which we will study two.

Before we start to introduce our methods, we make two adaptions to our raw data. First, we remove the baseline offset value of the samples. The offset value is dominated by the discrete baseline voltage output of the SQUID and if we do not remove it, the unsupervised methods mainly cluster events according to their discrete baseline offset. Those do not contain any useful information for our analysis. The second adaption does not affect the outputs of our method but decreases the run time significantly: We map our 8192-element time series $e_i[x]$ to 512-element time series $\tilde{e}_i[x]$, by building the average over tuples of 16 samples:

$$\tilde{e}_i[x] \leftarrow \frac{1}{8} \sum_{i=8x}^{8x+7} e[i]. \tag{5.1}$$

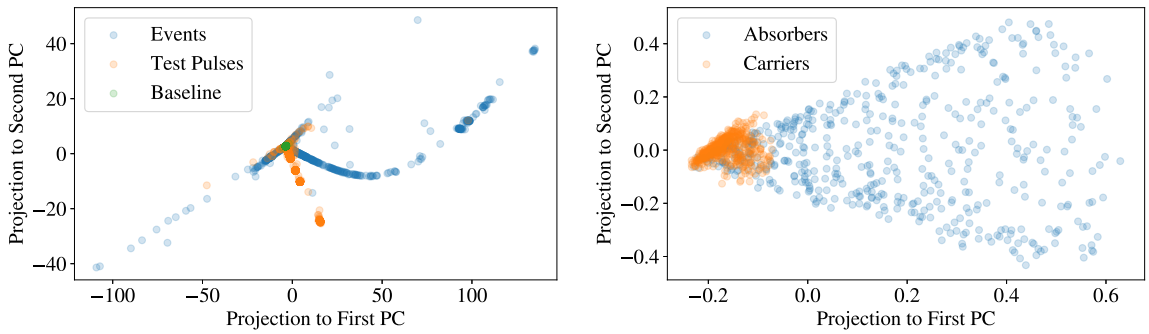
Here x denotes the array index. This preliminary reduction of dimensionality does only wipe out information about the background noise, which is of no interest to us anyway. We apply the same dimensionality reduction to our simulated absorber-carrier data set.

In the following section, we apply our first unsupervised machine learning method to the offset-removed, 512-dimensional raw data events and the simulated 512-dimensional carrier and absorber events.

5.2.1 Principal Component Analysis

The principal component analysis (PCA) is a standard method in statistics, widely used in the natural and social sciences for the exploratory analysis of data and for predictive models [77]. The idea is to find a basis in the high dimensional time series vector space, where the projection of our data to the first basis vector yields the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 5.2: (a) Projection of the raw data set to its first and second principal component. We see a good separation of the test pulses (orange) from the events (blue). The baselines (green) are intermixed with the low energetic events and test pulses. We can distinguish between two bands of events, from which one is a straight line, the other one is bent. The bending is most likely due to the saturation of high energetic events. (b) Projection of the absorber-carrier data set to its first and second principal component. The carrier events cluster around and intermix with the low energetic absorber events. Most likely the first two principal components superpose in a clever way such that they can describe the onset timing.

largest variance, the projection to the second basis vector the second largest, etcetera. These new basis vectors are called the *principal components* of the data distribution. We preserve as much variance as possible in our data when we project to a lower dimensional vector space, spanned by a number of principal components, starting from the lowest.

Mathematically the principal component analysis corresponds to an eigenvalue problem. We can write the whole data set as a data matrix $\mathbf{X}$, where each row contains the values of a sample and each column the value of a specific feature, in our case the corresponding element in the time series. Our data matrix has the shape 7000×512 for the raw data set and 10000×512 for the absorber-carrier data set. The first principal component is the vector

$$t_i = \mathbf{X} w_i, \tag{5.2}$$

where w_i is the eigenvector to the largest eigenvalue of the data matrix. The same applies to the higher principal components. We can therefore calculate the principal components by solving an eigenvalue problem, for which numerous optimized methods exist.

The PCA is a linear and invertible transformation, we can therefore transform a principal component back to our original 512 dimensional space. The back-transformed

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

first and second principal components of our raw data set both feature the typical pulse shape with only slight differences in the shape of their peaks. The projection of the raw data to its first and second principal component is shown in Fig. 5.2a. As the PCA is a linear transformation we expect events that feature the typical pulse shape, and have roughly the same onset timing, to build a band with the form of a straight line starting from the origin in the projection plot. We can see that the baselines (green) cluster in the origin of the two dimensional distribution, surrounded by low-energetic events (blue) and test pulses (orange). Starting from the origin, the test pulses build a band with dominantly high negative coefficients to the second component, well separated from other data points. The events do build a band with high positive coefficients to the first component. However, the band does not form a straight line, but a bent curve. This indicates, that the typical event shape changes with higher values – an effect that we believe is due to the saturation of high energetic pulses. In the last third of the band we can see a dense cluster of data points, indicating a high rate of pulses or artifacts with a specific height. We can identify a third, only weakly filled, band that stretches towards high negative values in the coefficients to both the first and the second component. This could probably be identified as a type of artifact, that occurs in different heights, but diverges from the typical pulse shape.

In Fig. 5.2b the projection of the absorber-carrier data set to its first two principal components is shown. The first and second component do again look like two pulse shapes, with slightly different form and onset timing. The absorber events are distributed in an longish triangular shape, the carrier events cluster around one edge of the absorber triangle. They do not separate well, we believe that the low energetic absorber events are intermixed with the carrier events. We also believe, that the first two principal components mainly describe the uniformly distributed time shift and do describe carrier events only as low energetic absorber events.

In summary, the principal component analysis succeeded in discriminating between test pulses and events in the raw data and helped to build a rough intuition about the distribution of artifacts. It did perform relatively well in distinguishing high energetic absorber and carrier events, but fails to do so for low energies. We did now only use the first and second two components. For practical applications, other than visualization, more components could be used. We have therefore reason to believe, that this can be an applicable method combined with an automatic clustering algorithm, i.e. Gaussian mixture models. Also, it can be used to compare the distribution of a simulated data set with the distribution of a measured, raw data set to find event types that are not yet included in the simulation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

There is a second unsupervised machine learning method that goes beyond a linear data transformation and projection and typically performs better in finding clusters in a given data distribution. We introduce it in the following section.

5.2.2 t-distributed Stochastic Neighbor Embedding

The t-distributed stochastic neighbor embedding (t-SNE) is currently the state-of-the-art technique for the visualization of data distributions [78]. Other than the PCA, that we studied in the previous chapter, it does not extract features from the samples, but solely measures the distance to their neighbors. It then constructs a lower dimensional distribution that resembles those neighbor distances as good as possible. The algorithm is somewhat involved and described in the following.

The underlying idea of the stochastic neighbor embedding is, that every sample in the high dimensional space chooses a neighbor stochastically. The symmetric probability that two samples x_i, x_j pick each other as neighbors is

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \tag{5.3}$$

where N is the size of the data set, p_{ii} is set to zero and $p_{j|i}$ is the conditional probability that x_i would pick x_j as its neighbor, based on their similarity. To describe this similarity, we exchange each sample with a Gaussian probability density $P_i(x_i, \sigma_i^2)$, centered at x_i and with bandwidth σ_i^2. The similarity $p_{j|i}$, which is a conditional probability, is defined as:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)}. \tag{5.4}$$

The bandwidth of the Gaussian kernel is chosen such that the *perplexity* of the distribution P_i equals a predefined value. The perplexity of P_i is given by

$$Perp(P_i) = 2^{H(P_i)}, \tag{5.5}$$

where $H(P_i)$ is the *Shannon entropy* of P_i:

$$H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i}. \tag{5.6}$$

The perplexity of P_i does only depend on the bandwidth σ_i, therefore the correct bandwidths can be found one after another. We do the same exchange of samples with probability densities for the lower dimensional space into which we want to map our data distribution. In a lower dimensional space a sample can have fewer neighbors,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

while maintaining the neighbors distances to each other. In order to compensate for this so-called crowing-problem, the Student-t distribution, which has heavier tails than the Gaussian distribution, is chosen instead of the Gauss distribution. By this, data points that have a moderate distance in the high dimensional distribution can be placed more flexible in the lower dimensional distribution, while close points stay close. The similarity between two samples y_i, y_j in the low dimensional distribution is therefore measured by

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}, \tag{5.7}$$

where q_{ii} is also set to zero. The data distribution in the higher dimensional space, i.e. all Gaussian kernels P_i together, is denoted as P, the distribution in the lower dimensional space as Q. The similarity of those distributions is optimized by a minimization of the *Kullback-Leibler divergence* of the distribution P from the distribution Q

$$\mathrm{KL}\,(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{5.8}$$

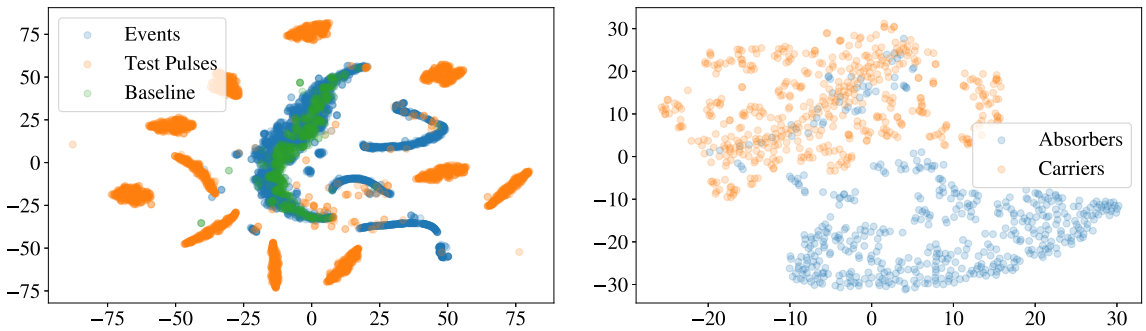
with respect to the position of the samples y_i in the lower dimensional space. The optimization is done with the gradient descent algorithm. The final outcome of the algorithm is a distribution of data points y_i in a lower dimensional space, that preserves distances between individual data points with special focus on the preservation of clusters.

There are two major differences compared to the PCA algorithm. First, the t-SNE is a nonlinear transformation. This means that a nonlinear feature in our events, as the onset time shift is, can be identified as only one feature. As we saw above, the PCA can only approximately describe the onset time shift as a superposition of multiple features. Second, t-SNE is a stochastic algorithm, i.e. it produces slightly different outcomes every time it is run. It is therefore common to run the algorithm multiple times until the outcomes are satisfying.

The perplexity parameter is typically chosen between 5 and 50 and is a smooth parameterization of the number of effective neighbors of each sample. However, the algorithm is very robust with respect to different perplexity values and we did only observe slight changes in the outcomes. After testing perplexity values between 5 and 100, we chose a perplexity of 50 for our final results.

The two dimensional visualizations we constructed with t-SNE for our raw data and absorber-carrier data sets are shown in Fig. 5.3. In the data distribution of the raw data set (Fig. 5.3a), we observe almost all test pulses (orange) to cluster according

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub



(a) (b)

Figure 5.3: Two dimensional data distributions, created with t-SNE. (a) The raw data distribution. The test pulses (orange) cluster according to their height. Four clusters of events (blue) distinct from the baselines are visible. In a large, longish cluster the baselines are intermixed with low energetic events. A higher density of baselines is visible on the right side of the cluster. (b) The absorber-carrier data distribution. The absorber events (blue) as well as the carrier events (orange) form distinct clusters, only low energetic absorber events intermix with the carrier cluster. We believe that the longish shape of the clusters is due to the onset timing of the pulses.

to their height. We see four clusters of events distinct from the baselines (green) and one large cluster of intermixed baselines and events. We believe that the large cluster represents low-energetic events that are more likely to neighbor empty baselines in the 512-dimensional sample space, while the distinct clusters represent events with higher energies. Within the large cluster a higher density of baselines is visible on the right side, supporting the hypothesis of intermixed low energetic events. The longish shape of the large cluster might be caused by variations in the baseline drift. However, a detailed description of the properties of the distinct clusters, especially which of them represent pulses and which artifacts, requires further studies, especially the per-hand analysis of several tens of events from each cluster. The data distribution of the absorber-carrier data set (Fig. 5.3a) shows a simple and clear structure: The carrier as well as the absorber events form a cluster. Some absorber events intermix with the carrier cluster, these are most likely events with very low energies. We believe, that the longish shape of the clusters is due to the onset time shift. However, the dimensions of a data distribution created with t-SNE do not correspond to any interpretable features of our data on a global scale, as the algorithm only tries to preserve distances.

Apparently t-SNE handles the scaling of events very well and builds compact clusters according to the overall shape of events, as long as these events occur with continuously distributed heights. This is a very useful property for our data, where artifacts and event types can occur with different amplitudes. For future applications

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the method should be tested within the region of interest, i.e. only for events with an energy below 40 keV. The clustering for low energetic events will certainly not be as clear as for the whole data set, but we believe that some meaningful structure should still be visible.

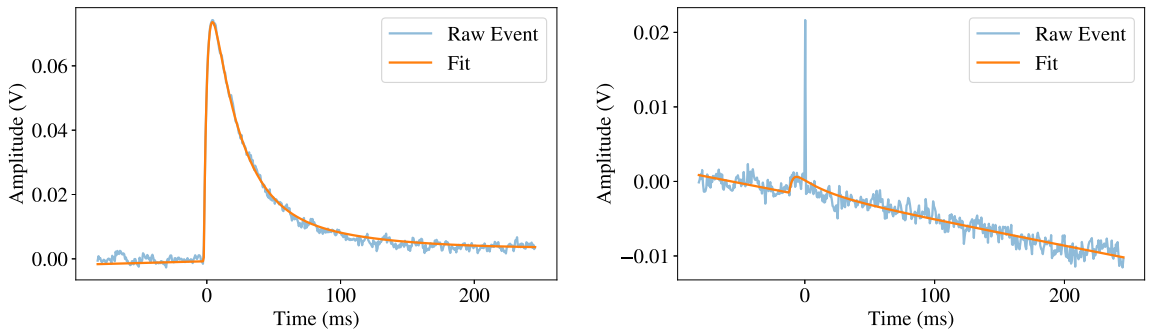
To summarize, PCA can provide a rough intuition about different types of events, while t-SNE succeeds in clustering event types together efficiently. There are manifold ways how this could be used in our analysis: First, the data distributions from different detectors can be visualized and compared. New types of events, that might only occur on specific detectors, can be identified. Both methods could be combined with an unsupervised classification algorithm, i.e. Gaussian mixture models, to provide a rough and automatic classification on new data from a not yet analyzed detector. For t-SNE a CUDA-compatible implementation exists, i.e. a version that uses the GPU instead of the CPU for calculations, which certainly makes it possible to analyze large data sets [79].

In the following section we introduce an algorithm for anomaly detection, which is also a class of unsupervised methods. It might be useful for the automatic detection of events that diverge from the typical absorber recoil pulse shape.

5.3 Anomaly Detection

In the past section we introduced two unsupervised methods with which we can explore our data distribution and possibly define general quality cuts to classify events as absorber recoils or artifacts. This task can also be done on an event-by-event basis rather than by analyzing the data distribution. The regular form of the time series that represent events are pulse shapes of various height, overlaid with noise. Everything that deviates from this form is an anomaly that we need to discard. A general quality cut for that purpose is the RMS-cut, which we introduce in the following section. In modern data science, for the task of anomaly detection often autoencoders are used. This is a specific neural network architecture that is trained to compress and reconstruct only regular events, such that the reconstruction error provides information about the regularity of an event. We are going to study autoencoders in the last section of this chapter.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 5.4: (a) Absorber recoil event (blue), fitted with the standard event (orange). The absolute fit error is 0.023, which indicates a good match for an event with a height of 70 mV. (b) Voltage spike (blue), fitted with the standard event (orange). The absolute fit error is 0.029, which indicates a deviation from the pulse shape for an event with height 20 mV.

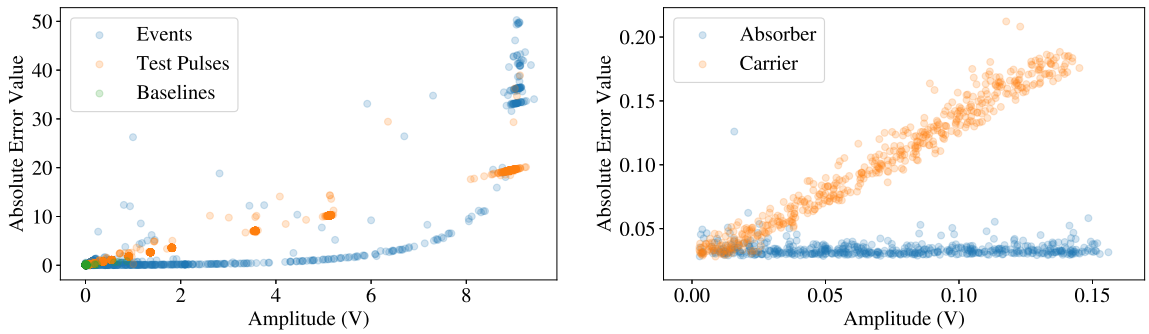
5.3.1 The Standard Event Fit

We already used the standard event fit in chapter 4 to reconstruct the height of events. Now we use it to gain information about the similarity between the shape of an event's voltage trace and the typical pulse shape. For this, we construct a standard event for absorber recoil pulses the same way we did in section 5.1 for carrier events and fit it with the pulse shape model. The fitted pulse shape model is superposed with a baseline offset and linear drift, as it was introduced in section 4.2. The combined pulse shape and baseline model can then be fitted to the raw data events and the absolute value of the fit error can be measured. We expect to measure a high error for events whose voltage trace shape deviates from the typical pulse shape.

In Fig. 5.4a the standard event fit (orange) to a raw absorber recoil event (blue) is shown. The model matches the raw event very well, a error value of 0.023 is measured. To provide a comparison, in Fig. 5.4b the standard event fit (orange) to an artifact, a voltage spike, (blue) is shown. The model cannot fit properly to the spike, which causes an absolute fit error of 0.03. This might not seem to be a striking difference, but the absolute fit error for absorber recoil events scales linearly with the pulse height. The absorber recoil pulse (Fig. 5.4a) is roughly three times as high as the voltage spike (Fig. 5.4b). Scaled accordingly, the absolute fit error for the voltage spike is therefore almost four times higher than that of the absorber recoil.

To determine which fit error is typical for regular and anomalous events, we apply the standard event fit to the whole raw and absorber-carrier data set. The resulting absolute fit error versus the height of the event are shown in Fig. 5.5. In the error plot

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

Figure 5.5: Standard event fit error versus pulse height (highest value in the time series of the event) for (a) the raw data set and (b) the absorber-carrier data set. In (a) we see several clusters of test pulses (orange) and for low pulse heights an almost constant or linearly with the pulse height error scaling for most events (blue). For events above 7 V, the error increases exponentially due to the saturation of the TES. The baselines are distributed close to the origin. In (b) we see an error scaling of the carrier events (orange) approximately proportional to the pulse height, while the error of the absorber events (blue) stays approximately constant.

for the raw data set (Fig. 5.5a) we can see a large band of events (blue) with almost constant error up to 4 V, which then increases almost linearly up to 7 V, where the saturation of pulses gets stronger. Above 7 V it increases roughly exponentially. This band most probably consists of regular recoil events, the fit error increases due to the saturation of the TES. The test pulses (orange) cluster according to their height and feature typically a higher fit error than the regular recoil events. The baselines (green) cluster around the origin. We can observe a large cluster of strongly saturated events with high fit error and pulse heights around 9 V. For the absorber-carrier data set (Fig. 5.5b), we see two clear, distinct bands, one for carrier events (orange) with linearly increasing fit error, and one for absorber events with almost constant fit error.

As we have seen, the standard event fit achieves a discrimination between regular recoil events and anomalous events. The different error scaling properties of different event types are especially interesting, which create distinct bands in the error plot. A discrimination between absorber and carrier events down to pulse heights of roughly 25 mV seems possible. We present now a second method for the detection of anomalies, namely autoencoding neural networks. These are already widely used in science and industry [80].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Model	Specifications	Learning Rate	Best Epoch	Best Validation Loss
ConvAE	2 Convolutional Layers, Kernel Size 3	0.01	16/20	**0.157736**
LSTM	3 Layers, Input Size 8, Hidden Size 4	0.01	25/26	0.165529

Table 5.1: Training results of an convolutional and an LSTM autoencoder. The convolutional autoencoder reaches a slightly lower validation loss value. The number of total epochs was chosen such that we did not expect a further decrease of the validation loss value for longer training. The learning rate was optimized on an trial and error basis.

5.3.2 Auto-Encoding Neural Networks

We already introduced neural networks in section 4.3 for regression tasks. An *autoencoder* is a stack of two regression networks, an encoding network $\mathcal{E}$ and a decoding network $\mathcal{D}$. The encoding network maps the vector that represents a sample x_i to a lower dimensional vector l_i, the latent state:

$$\mathcal{E}(x_i) = l_i. \tag{5.9}$$

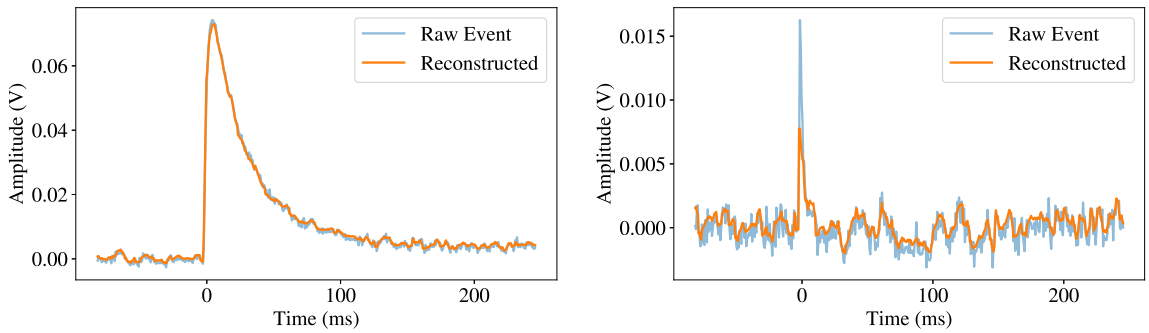
The decoding network maps the latent state back to an approximation of the original sample:

$$\mathcal{D}(l_i) \approx x_i. \tag{5.10}$$

The autoencoder, which is a stack of the decoder and encoder network, therefore learns to construct a compressed version of each sample, that still contains all information to reconstruct the sample. The training of an autoencoder needs no labels in the data set, it works completely unsupervised. The autoencoder can be used for anomaly detection, because it usually produces good results in reconstructing samples similar to the training set, but fails to reconstruct samples that have a different structure. The idea is to train the autoencoder on regular samples and then apply it to the whole data set. We expect a high error value for anomalous events, which helps to detect them.

We construct a data set of 9000 simulated TUM40 phonon events, with pulse heights uniformly distributed over the interval $[0\ \mathrm{V}, 10\ \mathrm{V}]$, which is the interval within that raw data events occur. We split the set into a training set of 7000 samples and a validation set of 2000 samples and train two different autoencoder architectures. The architectures resemble the corresponding architectures of regression networks, that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a) (b)

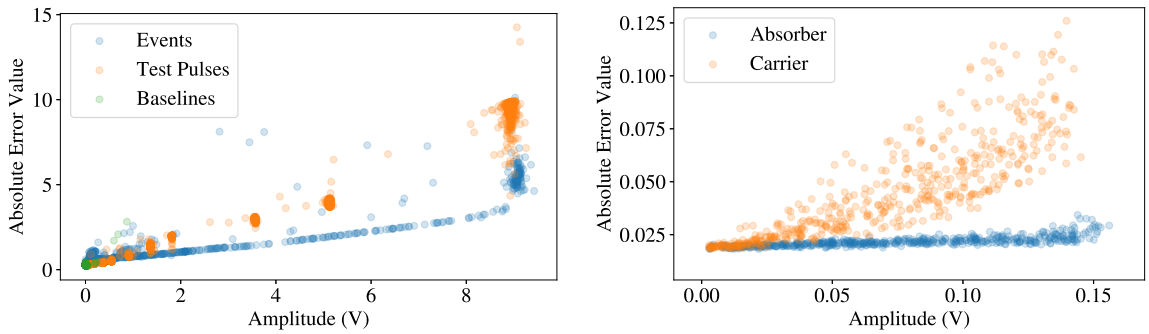
Figure 5.6: (a) Absorber recoil event (blue) and its reconstruction with an autoencoder (orange). The reconstruction error is 0.018. (b) Carrier recoil event (blue) and its reconstruction (orange) with an autoencoder. The reconstruction error is with an value of 0.02 relatively high for a 15 mV event.

we introduced in chapter 4 and for which we provide technical details in appendix A. The results of the training are listed in table 5.1 and are in favor of a convolutional autoencoder, a widely used architecture for samples in which elements are correlated with their neighbors. We therefore use the convolutional autoencoder architecture for this work.

In Fig. 5.6 a raw data absorber and a carrier recoil event are shown. The absorber recoil (Fig. 5.6a, blue) is very well reconstructed by the autoencoder (orange), causing a low reconstruction error of 0.018 for an event height of 60 mV. The carrier recoil (Fig. 5.6b, blue) deviates from its reconstruction (orange) due to its fast rising and decaying pulse shape. This causes a relatively high reconstruction error of 0.02 for an 15 mV high pulse. The autoencoder does not only reconstruct the pulse shape, but also many noise fluctuations. This overall error is therefore lower than for the standard event fit.

We apply the autoencoder to our raw data sets, the same way we did in the last section with the standard event fit, and record the reconstruction error versus the pulse height for each sample. We then repeat the process for the absorber-carrier data set, but this time the autoencoder is only trained on simulated absorber pulses with heights between 0 mV and 150 mV. This is the interval in which the pulse heights contained in the absorber-carrier data set are distributed. The results for both data set reconstructions are shown in Fig. 5.7. For the raw data set (Fig. 5.7a) we observe a similar structure as in the error plot of the standard event fit, even though the error value is overall lower. For the largest band, that consists of regular recoil events (blue), the error starts to increase linearly already for low pulse heights, for which is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) (b)

Figure 5.7: Reconstruction error versus pulse height (highest value in the time series of the event) for (a) the raw data set and (b) the absorber-carrier data set. In (a) we see a similar structure as in Fig. 5.5a, but with overall lower error and a slightly different error scaling of the largest band of events (blue). In (b) we see a significantly higher error for carrier events (orange) than for absorber events (blue), which makes a discrimination of events with regard to their reconstruction error possible.

stayed roughly constant for the standard even fit. The test pulse cluster (orange) are separated from other events, the baselines do also cluster around the origin. For pulse heights below 1 V the structure within the error plots look less clear than for higher pulses. The test pulse cluster seem to have lower error than the events, also the error does not scale linearly with the pulse height any more. For the absorber-carrier data set (Fig. 5.7b) we also observe a similar structure as we did in the error plot of the standard event fit (Fig. 5.5b). The absorber events (blue) have a roughly constant and very small error value, the error for the carrier events (orange) increases proportionally to the pulse height. Again, the overall error is smaller than in the standard event fit. An unsupervised discrimination of such high quality as we achieved it in the absorber-carrier data set is only possible because the absorber-carrier data set consists only of simulated events and the autoencoder was also trained on simulated absorber events. The training set was therefore very similar to the absorber events on that we evaluated the error values, leading to an unusually well reconstruction.

We have observed similar structures, but slightly different error scaling properties of regular recoil events in the autoencoder error plot (Fig. 5.7a) and the standard event fit error plot (Fig. 5.5a). We try now to combine there qualities by depicting the errors of both methods against each other within our region of interest, i.e. for pulse heights below $\sim$ 360 mV, which corresponds to recoil energies below 40 keV. The result is shown in Fig. 5.8. We observe clear and distinct clusters of test pulses (orange) and three distorted, band-like clusters of events (blue), out of which one cluster is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

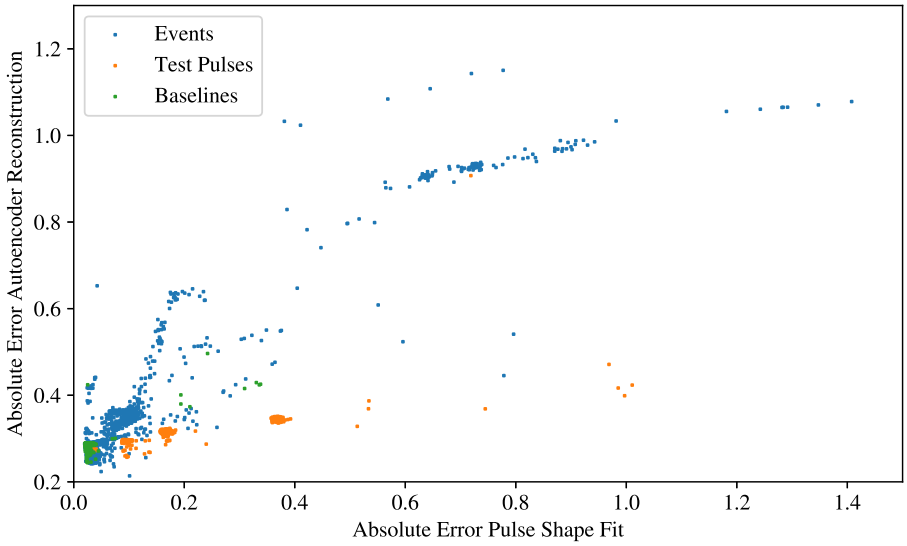


Figure 5.8: Autoencoder reconstruction error versus standard event fit error for raw data events within the region of interest, i.e. below 360mV/40keV. We see three distinct clusters of test pulses (orange) and three distorted, band-like clusters of events (blue), together with one cluster of intermixed events and baselines (green), closest to the origin.

intermixed with baselines (green). A decent analysis of the event clusters could bring insights about the structure of regular and anomalous events in our data. We believe that it is possible to identify individual clusters or bands of each type of events in this cross-error plot (Fig. 5.8). The pulse height of events could be incorporated with colors, as well as different properties, e.g. from section 5.2. We believe that the autoencoders error scaling for low energetic events would become more linear if we train the autoencoder only on absorber recoil events with energies within the region of interest. Then, individual bands within the error plot corresponding to specific event types could be identified even better.

We introduced in this chapter two methods to reduce the dimensionality of a data distribution and two methods to separate regular and anomalous events. We could identify a typical structure within our raw data: Typical clusters of test pulses and between two and four distinct bands or clusters of events were visible, together with one large cluster of intermixed baselines and low energetic events. For future applications, general quality cuts could be defined based on combinations of these unsupervised methods. They also introduce a way to visualize the data of just one detector, which was until now only possible for a detector module, by the light yield plot.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

It is in the nature of unsupervised classification methods that obtained results are less precise than those from supervised classification methods. In the last chapter we study deep, supervised classification with the classes of simulated events that we introduced already. These could be used to define specific quality cuts, i.e. to identify specific and well-known artifacts in our raw data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Chapter 6

Supervised Event Classification

If there is one machine learning technique that has changed the world in the past couple of years, it is supervised classification with neural networks. All modern image and voice recognition algorithms are based on this technology and it is applied in every smartphone, by every online social network, all autonomously moving vehicles, and many more [81–83]. Supervised classification does indeed produce incredible results – but there is one catch. The quality of all results is maximally as good as the quality of the data, on which the neural networks were trained. For this reason, the assembly of labeled real-world data has become the gold rush of the twenty-first century. However, often there is barely a way around labeling data by hand, which can become a cost-intensive task, depending on the complexity and diversity of the data. For supervised classification with neural networks a data set with a size of typically several hundred to thousand samples per event class is needed. For diverse data, as images for object identification, the data set has to be accordingly larger.

Supervised classification could be used as a specific quality cut in our analysis framework, discriminating classes of artifacts and regular absorber recoil events. So far we studied three different classes of events in detail: Recoil events within the absorber and carrier crystals, and empty noise baselines. We overcome the need for a labeled, raw data set by using simulated and automatically labeled data. We introduced the methods for the simulation of the above mentioned event classes in the chapters 3,4 and 5. The simulation was modeled from data of the TUM40 phonon detector, that was used in run 33. In the first section of this chapter, we train a neural network to discriminate between absorber and carrier recoils, in the second section we apply the same network architecture to discriminate between empty noise baselines and regular recoil events.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Model	Specifications	Learning Rate	Best Epoch	Best Validation Loss	Accuracy
FNN	2 Layers, 1000 Hidden Nodes	0.001	1/7	0.068	0.932
CNN	2 Convolutional Layers, Kernel Size 8	0.001	14/16	0.004664	**0.994**
ResNet18	-	0.001	9/11	0.007481	0.989
LSTM	3 Layers, Input Size 8, Hidden Size 80	0.001	12/20	**0.004527**	**0.994**

Table 6.1: Validation loss values, accuracies and training specifications for the feed forward net, convolutional net, ResNet18 and LSTM net. The networks were trained to discriminate between simulated absorber and carrier recoils. We see the highest accuracies and lowest validation loss values for the convolutional and LSTM networks. All hyperparameters were optimized on a trail-and-error basis.

6.1 Classification of Carrier Recoils

The carrier event was introduced in section 5.1. It originates from a recoil within the carrier crystal, a specific property of the composite design. The carrier event has to be filtered with a specific quality cut, because its pulse shape deviates strongly from regular absorber recoils, which makes an energy calibration impossible. In the current CRESST analysis framework, a special variant of the standard event fit is used for this purpose: The standard event of the absorber and the carrier recoil event are fitted to the raw data events and both fit errors are recorded. Their difference is normalized to a value in the interval $[-1, 1]$, events above a certain threshold are discarded. This method is explained in detail in Ref. [31] and was applied in the analysis of the TUM40 phonon detector from run 33. We see in Fig 2.5 the efficiency, i.e. the survival rate of absorber recoil events, of this special quality cut, the carrier cut (blue line). We follow a different approach in this work, we want to discriminate between absorber and carrier recoils with neural networks. Deep learning was already used in Ref. [14] for this task, we will compare their results with ours shortly, after we have introduced our own neural network model.

A neural network for classification follows the same working principle as a network for regression (see section 4.3). We design our network for binary classification, i.e. discrimination between only two event classes. We chose the last layer of our network with only one node, the label equals 1 for absorber recoil events and 0 for carrier recoil events. For predictions, we impose a decision rule to the output of our

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

network. The common decision rule, that we use as well, is to count predictions with a value above 0.5 as absorber events and below 0.5 as carrier events. If we use our network as a specific quality cut, the prediction of an absorber event would correspond to the survival of this event, while the prediction of a carrier event would lead to the discarding of it. Commonly the cross-entropy loss function is used for classification tasks. However, for our work we use the mean square error loss function instead, because we do not want to emphasize samples that are difficult to predict, i.e. events with very low energies, at the cost of all other samples in the training process. This strategy was proposed in Ref. [84].

For the training of our neural network we create a data set of 4500 absorber and 4500 carrier recoil events with heights uniformly distributed in the interval $[0\ \mathrm{mV}, 150\ \mathrm{mV}]$ and onset timings uniformly distributed in the interval $[-20\ \mathrm{ms}, 20\ \mathrm{ms}]$. We train four different neural network architectures for the discrimination task: A feed forward network, a convolutional network, a ResNet18 and an LSTM network. We already introduced the latter three architectures in chapter 4. The feed forward network, also called *dense* neural network, is the most basic version of a neural net that we introduced in the beginning of section 4.3. The data set is split into a training and a validation set with the ratio 7 : 2. The resulting validation loss values and the specifications of the training process are summarized for all four network architectures in table 6.1. All hyperparameters, i.e. number of nodes and layers and the learning rate, are optimized on a trial-and-error basis. The *accuracy* is the score of true predictions divided by the number of total predictions. We see overall accuracies above 0.9, indicating that all networks are capable of discriminating between absorber and carrier events. However, the convolutional and LSTM network perform significantly better than their competitors with accuracies of 0.994. The LSTM network has a slightly lower validation error, for which reason we chose this architecture for our discrimination task.

To test the performance of the LSTM network, we use a test set that consists of 5000 absorber and 5000 carrier events with heights between 0 and 15 mV. For the evaluation, we focus only on the most challenging regime of low recoil energies. The predictions are binned in intervals of 0.5 mV and the averaged predictions within each bin are shown in Fig. 6.1, for absorber and carrier events separately. These are the survival rates of the absorber (blue) and the carrier (orange) events after the application of the specific LSTM carrier cut. We can see a survival rate for absorber events above 0.8 for pulse heights of roughly 3.5 mV and above. The survival rate of carrier events is never above 0.2 and decreses below 0.1 for pulse heights above 6 mV.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

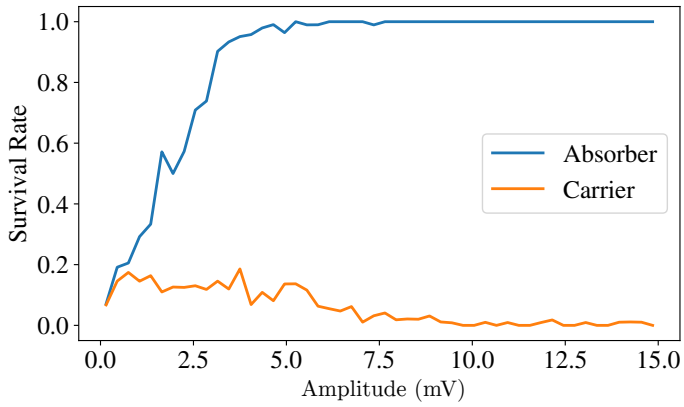


Figure 6.1: Survival rate of absorber (blue) and carrier (orange) events, discriminated by an LSTM neural network. We see that the survival rate of the absorber events increases to a value above 0.8 for pulse heights above roughly 3.5 mV.

The results, depicted in Fig 6.1, can be compared to the efficiencies from the run 33 analysis of TUM40, depicted in Fig. 2.5. The carrier cut (blue line) reaches a survival rate of simulated absorber recoil events of 0.8 at roughly 18 mV. This is a factor 5 higher than the value we achieve with the LSTM carrier cut. Considering that the ability to apply quality cuts on low energetic events is necessary to perform a meaningful search for low energetic dark matter, this improved discrimination threshold could add significant value to the analysis of detectors, which use the composite design. A carrier cut based on convolutional neural networks was already studied in Ref. [14]. We believe that the results of their work are of equivalent quality to ours. They succeed in discriminating carrier and absorber events above a pulse height of 10 mV with an accuracy of almost 1, same as we do. However, they do not provide a survival or efficiency plot or discrimination threshold, to which we could compare results directly.

The LSTM carrier cut is a proof of principle, that neural networks can achieve good results for the classification of raw data events. Every artifact that can be filtered within the region of interest directly improves the calculated dark matter exclusion limit and therefore the overall value of CRESST's experimental results. We repeat this procedure for the discrimination between noise baselines and low energetic absorber recoils in the following section.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Model	Specifications	Learning Rate	Best Epoch	Best Validation Loss	Accuracy
FNN	2 Layers, 1000 Hidden Nodes	0.001	5/10	0.021002	0.979
CNN	2 Convolutional Layers, Kernel Size 8	0.001	18/20	0.004863	**0.995**
ResNet18	-	0.001	11/15	0.017905	0.981
LSTM	3 Layers, Input Size 8, Hidden Size 80	0.001	20/20	**0.004339**	**0.995**

Table 6.2: Validation loss values, accucies, training specifications and hyperparameters for all four neural network architectures that were trained to discriminate between absorber and carrier recoils. We observe the highest accuracy for the convolutional and the LSTM network, while the LSTM network achieves a slightly lower validation loss value.

6.2 Classification of Noise Baselines

The trigger threshold in the current CRESST analysis framework is determined as follows: A large number of measured noise baselines is run through the analysis chain and the so reconstructed height values of surviving events are recorded. The trigger threshold is then chosen such that only a small number of surviving noise baselines is above the threshold. These surviving baselines above the threshold are wrongly counted as low energetic events. By this, a trade-off between a low trigger threshold and a certain number of noise triggers is made. A precise discrimination between noise baselines and low energetic recoil events could filter wrongly triggered events and therefore lead to an improvement of the dark matter exclusion limit.

We want to accomplish this discrimination with a neural network model, as we did in the previous chapter. In this binary classification task the labels are chosen such that 0 corresponds to a noise baseline and 1 to an absorber recoil event. For the training of our neural network we simulate a data set of 4500 absorber recoil events with pulse heights uniformly distributed in the interval $[0\ \mathrm{mV}, 150\ \mathrm{mV}]$ and onset timings uniformly distributed in the interval $[-20\ \mathrm{ms}, 20\ \mathrm{ms}]$, and 4500 empty noise baselines. The data set is split into a training set and a validation set by the ratio of $7:2$. The events, we simulated, were constructed with the same methods which we introduced in chapter 3. The same neural network architectures as in the previous section are tested: A dense neural network, a convolutional neural network, a ResNet18 and an LSTM. The validation loss values, accuracies and specifications

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

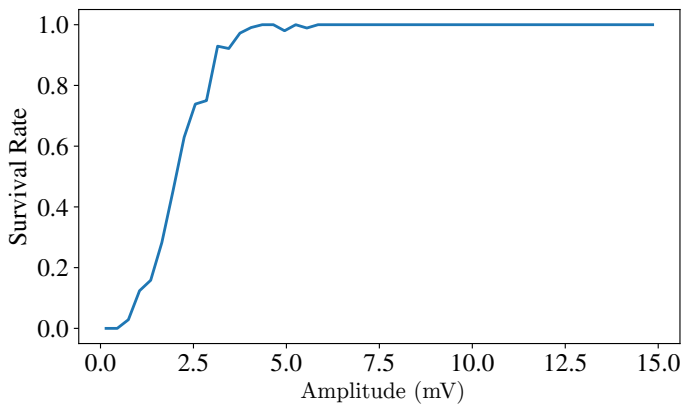


Figure 6.2: Survival rate of absorber recoil events, i.e. rate of correctly classified absorber recoil events, after a classification with the LSTM noise cut. We observe a rise of the survival rate above 0.8 at 3.5 mV. The survival rate of noise baselines, i.e. rate of wrongly classified noise baselines, is 0.00154, which corresponds to roughly one out of 650 baselines.

of the training process and hyperparameters are listed in table 6.2. We see roughly similar results as in the previous section: The convolutional net and the LSTM model yield the best performance with an accuracy of 0.995, the LSTM has a slightly lower validation loss value than the convolutional network. We therefore use the LSTM model for our work.

We test the model on a set of 5000 noise baselines and 5000 absorber recoil events with heights distributed uniformly in the interval $[0\ \text{mV}, 15\ \text{mV}]$ and onset timings as in the training set. The predictions for absorber recoils are binned in intervals of 0.5 mV, the average value of each bin is depicted in Fig. 6.2. This is also the survival rate of the absorber recoil events, when we apply the LSTM model as a specific quality cut. We see a survival rate of 0.8 for pulse heights of roughly 3.5 mV and above. The survival rate of the baselines is 0.00154, which corresponds to one out of 650. This indicates that the LSTM model performs well in identifying noise baselines, but counts some low energetic events accidentally as noise as well.

The currently lowest detection threshold for WIMP events from the CRESST-III experiment was achieved with detector A in run 34. This detector is roughly 10 times smaller than the TUM40 detector and has a conversion rate from measured pulse height to recoil energy of only 4.38 keV/V, while that of TUM40 is 111 keV/V, far higher. We can therefore not directly compare the events of those two detectors, however, we want to study the implications that an LSTM noise cut would have to the detection threshold anyway. The threshold for detector A was chosen such that only one noise trigger occurs within an exposure amount of 1 kg days. The relation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

between the threshold and the number of noise triggers is depicted in Fig. 6.3. The

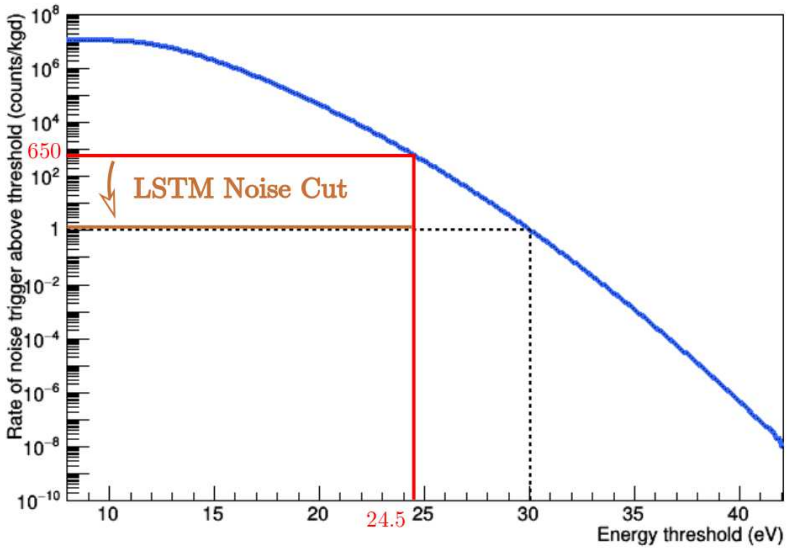


Figure 6.3: Relation between trigger threshold and rate of noise triggers (blue line) for detector A, used in run 34. The threshold (30.1 eV) was chosen, such that only one noise event is triggered per kg days exposure (black dotted line). In our evaluation, only one out of 650 noise baselines survived our LSTM noise cut. The threshold could therefore potentially be lowered to 24.5 eV, which would cause 650 noise triggers per kg days exposure. After the application of the LSTM noise cut, only one noise trigger would survive, which was the initial rate. Picture adapted from Ref. [8].

constraint on the number of noise triggers determines the threshold of 30.1 eV (black dotted line). Assuming that the LSTM noise cut works roughly as well on data from detector A, as it does on data from the TUM40 detector, we could allow for 650 noise triggers per kg days. The LSTM carrier cut reduced the noise trigger rate back to its initial value of one (brown line), while the low energetic recoil events would survive. By this, the threshold could be lowered to a value of 24.5 eV (red line), without further adaptions to the analysis framework. Even though we have to train and evaluate the LSTM noise cut on data from detector A before we can be sure about its implications for our current dark matter limit, we believe that an improvement of the threshold compatible to the one shown in Fig. 6.3 could be possible.

To summarize, we trained two LSTM neural networks to perform specific quality cuts on simulated raw data events. The first model discriminates reliably between carrier and absorber recoil events with a pulse height greater than 3.5 mV, which is a 5 times lower threshold than it was achieved in the analysis of the TUM40 detector in run 33. The second model discriminates reliably between empty noise baselines

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and absorber recoil events with roughly the same pulse height threshold, which could potentially reduce CRESST's currently lowest detection threshold from 30.1 eV to 24.5 eV.

However, all those results were only achieved on simulated data. The application to real, measured data is more complex and requires a careful evaluation. In the last section of this chapter, we evaluate our two models on a small sample of measured and hand-labeled data and provide a rough intuition of the potential and the challenges that might come with the application of neural networks in the analysis of the data from a whole CRESST run.

6.3 Application to Measured Data

In the previous sections we achieved good results for the supervised classification with neural networks on simulated absorber and carrier recoil events and simulated noise baselines. The ultimate goal of all the methods we develop has to be a good and robust performance on real, measured data. The evaluation on measured data is difficult, because we lack reliable labels for the events. The data acquisition labels automatically recorded noise baselines, test pulses and triggered events (see section 2.3). However, roughly 2% of recorded noise baselines contain accidentally recorded events and only a fraction of all triggered events are recoil events. The larger fraction consists of temperature fluctuations, offset jumps of the SQUID-based readout system or voltage spikes.

To provide a reliable evaluation, we chose a small sample of 200 raw data events for the TUM40 phonon detector and labeled them by hand. To each triggered event we assigned one of the following three classes: Absorber recoil, carrier recoil or artifact. The baselines and test pulses were automatically labeled by the data acquisition, although we checked that no artifacts are visible on the baselines, which was not the case. The resulting counts for each event class are documented in table 6.3. We see that the largest class of triggered events are artifacts, followed by carrier recoils. Only 19 out of the 88 triggered events are absorber recoils. The output values of the LSTM carrier cut for the triggered events versus the event height in Volt are depicted in Fig. 6.4a. The LSTM carrier cut was only trained on events with pulse heights up to 150 mV, therefore we only show predictions for events up to this pulse height. We can observe that all carrier (red) and absorber (cyan) recoils are correctly classified, while the artifacts (green) are classified as absorbers or carriers alike. Only 6 absorber events are visible within the interpolation region, i.e. up to 150 mV. From the remaining

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

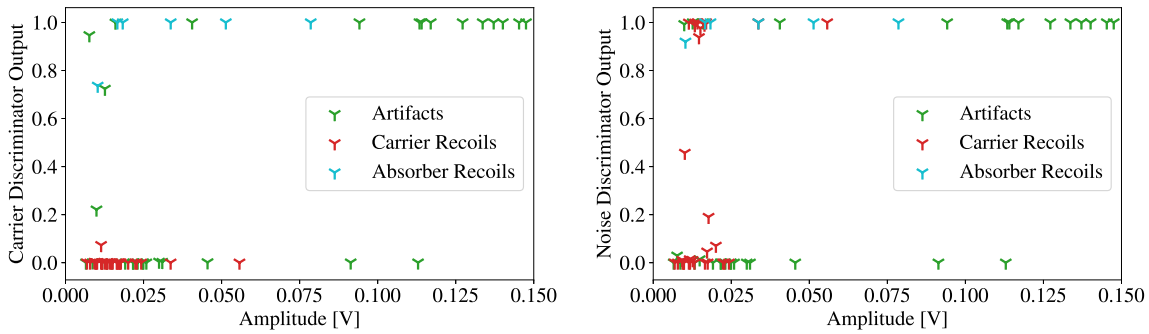
Event Type	Counts	Misclassified Events by LSTM Noise Cut	Misclassified Events by LSTM Carrier Cut	Remark
Absorber Recoils	19	0	11	All misclassified events are outside the interpolation interval.
Carrier Recoils	31	-	0	-
Artifacts	38	-	-	Mainly Temperature Drifts, Decaying Baselines.
Baselines	18	3	-	Automatically labeled.
Test Pulses	94	-	-	Automatically labeled.

Table 6.3: Counts of event classes and misclassified events by the LSTM carrier and noise cut for a hand-labeled data set of 200 measured events from the TUM40 phonon detector. The baselines and test pulses are automatically labeled by the data acquisition, the baselines do not contain any accidentally recorded artifacts. All absorber recoils that were classified as carrier recoils have pulse heights above 150 mV, which is the highest pulse height in the LSTM carrier cut's training set.

13 absorber recoils, 11 are misclassified as carrier events. This does not come as a surprise, as neural networks can typically only interpolate between the training data, but not extrapolate to samples that feature different properties. For the same reason the predictions of both our LSTM models on artifacts do not contain any value for our evaluation, as no such artifacts were in our training set.

The output values of the LSTM noise cut for the triggered events are shown in Fig. 6.4b. All absorber recoils (cyan) are correctly classified, the carrier recoils (red) and artifacts (green) are classified as events and noise baselines alike. This is again not surprising, as the training set of the LSTM noise model did only contain absorber recoils and noise baselines. In Fig. 6.5, the outputs of the LSTM noise cut on the recorded noise baselines are shown. We observe three out of 18 baselines that are wrongly classified as events. These baselines have a relatively strong positive drift in common, which leads to a high maximal value of the noise fluctuations. However, the simulated training and validation set did contain noise baselines with even higher maximal values and stronger drifts, that were correctly classified as noise baselines. We believe that the three wrongly predicted baselines differ from the baselines in the simulated data set in some other property, which should be investigated in detail for future applications. However, the output value for these three baselines is around 0.99,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) (b)

Figure 6.4: (a) Output values of (a) the LSTM carrier cut and (b) the LSTM noise cut for 88 triggered and hand-labeled events. The output value one corresponds to an absorber recoil, the value zero to a carrier recoil or noise baseline. (a) We see that all absorber (cyan) and carrier (red) recoil events are correctly classified. No artifacts were in the training set, therefore artifacts (green) are unpredictably classified as absorber or carrier recoils. (b) Also here all absorber recoils (cyan) are correctly classified, while the carrier events (red) and artifacts (green), which event classes were not included in the training set, lead to unpredictable outputs.

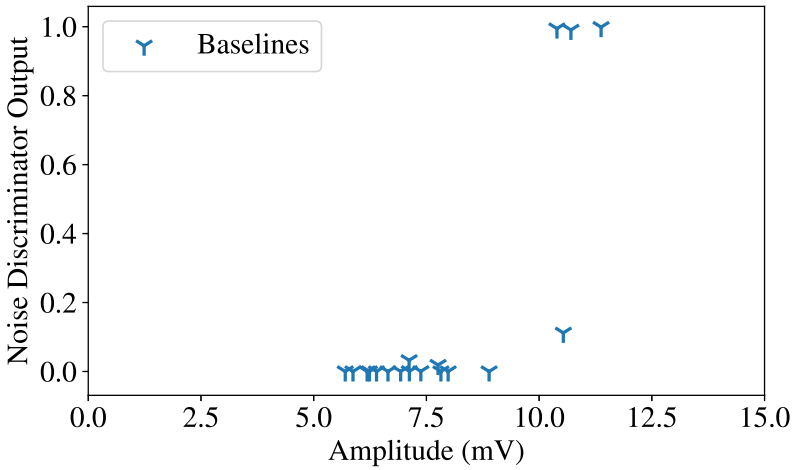


Figure 6.5: Output values of the LSTM noise cut for the triggered events from a hand-labeled data set of size 200. The output value one corresponds to an absorber recoil, the the value zero to a noise baseline. Three out of the 18 baselines have values close to one, which would correspond to a wrong classification if we use the typical cutoff value of 0.5.

while the output value for the absorber recoil events equals exactly one for all events but one low energetic event (see upper-left edge of Fig. 6.4b). We could therefore make a trade between three wrongly classified baselines and one correctly classified low energetic absorber recoil by adapting the cutoff value of the decision rule for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

predictions accordingly.

As we have seen, there are differences between our simulated data and the measured data. Both LSTM-based cuts work very well on events and baselines that have similar properties as the simulated training set has. The classification of artifacts is not predictable and we identified some potential shortcomings in our noise baseline simulation. For future applications a medium-sized data set of measured events should be assembled and labeled by hand. Also typical artifacts should be simulated and the neural network model should be modified to perform multiclass classification, instead of only binary classification. A system to check the reliability of predictions of the classification network would also add value to our method, this could be done with ensemble learning, Bayesian networks or Monte Carlo dropout [74, 85, 86]. Currently a lot of research is invested in self supervised algorithms, i.e. algorithms that start with a small data set of labeled data and attach labels to originally unlabeled data automatically throughout the training process [87]. Models of this type could be applicable to our type of data and we should monitor the developments in this field.

After all, we are confident that supervised classification with neural networks will show reliable, high-quality results as specific quality cuts in the CRESST analysis in the long run.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Conclusion

In the first two chapters of this thesis we introduced the experimental evidence for the existence of dark matter, suitable dark matter candidates, as well as the working principle, detector concept and data analysis framework of the CRESST experiment.

In the third chapter we presented a method for the simulation of recoil events and noise baselines, modeled after data from the TUM40 phonon detector, and studied the saturation effect of the detectors. We investigated a possible, yet unknown, non-linear energy loss of high energetic pulses, that might have an impact on the truncated fit, a common technique in the analysis of CRESST data. We have shown that this energy loss might be due to self-heating and have a stronger impact on detectors that work as calorimeters.

In the fourth chapter we compared the reconstruction of pulse heights and onset timings of simulated recoil events with the optimum filter, the standard event fit and an LSTM model. We achieve the best height reconstruction for low energetic events with the LSTM model, that is capable of reconstructing pulse heights between 0.6 mV and 6.8 mV with a bias and standard deviation of less than 0.1 mV each. The ability to reconstruct the energy of recoil events with a height of 0.6 mV for the TUM40 detector, which is a CRESST-II detector, corresponds to a detection threshold of 67 eV, typical for CRESST-III detectors. The best reconstruction of the onset timing for pulse heights was achieved with the standard event fit. All methods were exemplarily applied to raw data events and provided comparable outputs.

In the fifth chapter we introduced a simulation method for the carrier event, a typical artifact of the TUM40 phonon detector. We then analyzed the raw data distribution and the data distribution of simulated recoil events with two unsupervised techniques for dimensionality reduction: The principal component analysis (PCA) and the t-distributed stochastic neighbor embedding (t-SNE). Especially the t-SNE provided a clear and distinct clustering of different types of events. We therefore found a suitable method for the visualization of data from a single detector, which was before only possible for two detectors of a module combined, with the light yield

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

plot. In the last section, we applied the standard event fit and autoencoders to detect anomalous events, i.e. events that deviate from the typical pulse shape. It was possible to identify different types of potentially anomalous event classes with a combination of both methods.

In the sixth chapter, we trained two LSTM models to discriminate between simulated carrier and absorber recoils, as well as between simulated recoil events and noise baselines. The models achieve a discrimination threshold of roughly 3.5 mV, which is a third lower than the TUM40 detection threshold in the run 33 analysis. We discussed a potential lowering of the currently lowest CRESST-III detection threshold from 30.1 eV to 24.5 eV, by the use of the LSTM noise discrimination model. Both LSTM models are evaluated on a small sample of raw, hand-labeled TUM40 phonon events. The LSTM carrier discriminator succeeds in classifying all absorber and carrier events correctly, while the LSTM noise discriminator classifies three out of 18 baselines wrongly as recoil events, indicating possible shortcomings in our simulated noise baselines.

Future work could build upon our results in various directions. The simulation of saturated events could be accomplished by the inclusion of the self-heating effect and the superconductor's transition curve in the derivation of the pulse shape model. This would most probably not yield an analytical fit function, however, the equations describing the temperature within the TES could still be solved numerically. For the transfer of our deep learning methods for simulated events to the raw data analysis, a medium-sized and hand-labeled data set of events from different detectors should be assembled. Also, simulations of typical artifacts should be included in the simulated data set. With these data sets, neural networks for the classification of multiple event classes can be trained, which can then serve as specific quality cuts. A software for the convenient visualization and exploratory analysis of raw data could be based on the methods for dimensionality reduction and anomaly detection. The latter ones could also be used to define general quality cuts. A clever combination of multiple machine-learning based and knowledge-based methods for the height reconstruction of events could lead to a lowering of the trigger threshold and provide feedback about the accountability of the predicted values. Another set of machine learning methods, that were not investigated in this work, are reinforcement learning methods, typically used for decision making processes of active agents. The simulation methods we introduced in our work could be used to build a simulated environment for a reinforcement learning agent, that could learn to control the bias and heater currents of all detectors

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

and by that optimize their signal-to-noise ratio and life time, i.e. time with stable operation point.

All that said, we are convinced that machine learning methods will greatly improve the experimental results of CRESST and similar cryogenic experiments.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Appendix A

An Introduction to Deep Learning

We typically differentiate between three types of machine learning: *Supervised learning*, *unsupervised learning* and *reinforcement learning*. We already introduced the basic ideas of supervised learning at the beginning of chapter 3 and those of unsupervised learning in chapter 4. We do not consider reinforcement learning in the work at hand, which is usually applied for decision making processes of active agents. A good introduction to reinforcement learning can be found in Ref. [88]. In section 4.3, we introduced the basics of *deep learning*, i.e. machine learning with neural networks. We summarize all concepts of supervised learning in the following and give a more detailed explanation of the neural network architectures we use. We do not repeat the concepts of unsupervised learning in this appendix and refer for that to chapter 5 and Ref. [69].

A supervised machine learning method is a *function approximator*, i.e. for a given data set of *samples* x_i and corresponding *labels* y_i, it learns a function $\hat{f}$ that assigns to each sample a prediction of its corresponding label:

$$\hat{f}(x_i) \approx y_i. \tag{A.1}$$

The labels can be continuous numerical values, which corresponds to a *regression problem*. They can also be categories, which are typically represented by one binary variable for each possible category and label. This is called *one-hot encoding* [89]. The prediction of categories would correspond to a *classification problem*. For classification problems, a decision rule g is applied to the continuous output of the function approximator $\hat{f}$, that maps the continuous values to the categorical label:

$$g(\hat{f}(x_i)) \approx y_i^{(categorical)}. \tag{A.2}$$

Usually this decision rule chooses the corresponding category of the highest output value of $\hat{f}$. In our work, we solve only binary classification problems, the output of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the function approximator is therefore a value between zero and one, where the exact values zero and one correspond to the two categories. The decision rule, a cutoff value is chosen, above which the continuous output is rounded up and below that it is rounded down. The typical cutoff value is 0.5.

Every supervised learning model is *trained*, i.e. fitted/optimized, for the best possible *predictions*, i.e. approximations of the label, for a set of training data. This is typically done in an iterative optimization process, throughout which the performance on the training data set is compared with the performance on a *validation* data set. By this, so-called *overfitting* is prevented, which would mean that the model gives good predictions on the training data, but does not generalize well to other data, on which it is applied later on. Final tests of the performance are done on a third data set, the *test set*. While the training and validation set are typically randomly split, the test set is an assembly of representative samples, to reach meaningful statements about the performance of the model in applications.

Neural networks are a type of supervised learning models, that perform especially well on high dimensional and highly correlated data [90]. We introduce the various architectures, that we use in our work, in the following. In the last two sections of this chapter, we introduce two techniques that use neural networks also for unsupervised learning tasks, that we applied in chapter 5 and appendix B.

A.1 Feed Forward Networks

We introduced the basic principles of machine learning in the beginning of this appendix and start now with the basics of deep learning, i.e. machine learning with neural networks. The most simple form of a neural network $\mathcal{F}$, called *feed-forward (FNN)*, *dense*, *fully connected* or *vanilla*, consists of at least two matrices L_0, L_1, two bias vectors b_0, b_1, filled with *weights*, and a non-linear *activation function* σ:

$$\mathcal{F}(\cdot) \equiv L_1(\sigma(L_0(\cdot) + b_0)) + b_1. \tag{A.3}$$

The weights are fit coefficients that are trained during the training process and determine the quality of the predictions of the network. The activation function is a piece wise smooth and non-linear function, typically with a Lipschitz constant smaller than one. A common choice for the activation function is either the sigmoid function or the *ReLU* function, that outputs zero for negative input values and is the identity function for positive input values [91].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

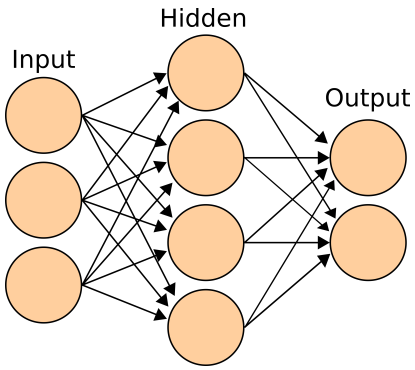


Figure A.1: Architecture of a minimal simple feed forward neural network. The nodes are structured in one input layer, one output layer and one hidden layer, which are interconnected by the weights of the network [92].

This architecture is depicted in Fig. A.1. The input values of the network are typically called the *input nodes*, depicted on the left side of Fig. A.1. The set of all input nodes is called the *input layer*. The output values of the activation function are called the *hidden nodes*, together they form the *hidden layer*. The same principle applies to the output values of the network, which are called the *output nodes* and constitute the *output layer*. The weight matrices are used to compute weighted sums from the nodes in the previous layer for every node in the subsequent layer. Not explicitly shown in Fig. A.1 are the bias values, which simply add a trained and constant value to every node. The number of input and output nodes are chosen according to the dimensionality of the samples and labels, the number of nodes in the hidden layer is typically optimized according to the complexity of the approximation problem.

Neural networks are trained with adaptions of the *gradient descent* algorithm [93] and the *backpropagation* algorithm [94]. This is an iterative process, each step of the optimization is called an *epoch*. In each epoch, the network predicts values for each sample from the training set and the deviation of the predictions from the labels is measured by a *loss function*. For regression problems, the mean squared error function is often used, for classification problems the cross-entropy loss function [84]. The weights of the network are then updated towards a better match of the predictions with the labels. The most important parameter for the training process is the *learning rate*, that controls the magnitude of each weight update. Typically weights are not updated for the calculated loss value of the whole training set at once, but for *mini batches* with size between 8 and 128 samples. Parameters of the training process, as the learning rate and the batch size, and the network architecture, as the number of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

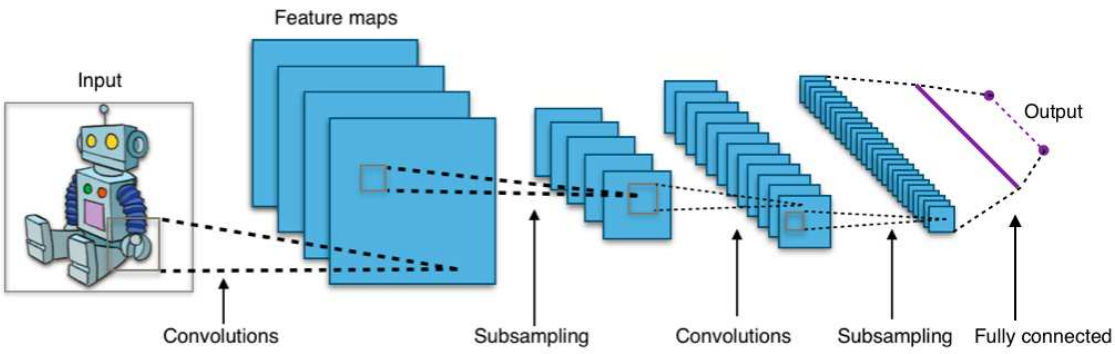


Figure A.2: Typical architecture of a convolutional neural network. The input sample (a picture of a robot in this example) is convoluted with weight kernels, the result of the convolution is called feature map. The dimensionality of the feature map is decreased by subsampling and a activation function is applied. The combination of the convolution, subsampling and activation function is called a convolutional layer. Typically many convolutional layers are stacked on top of each other, before the outputs of the last convolutional layer are inserted into a fully connected layer, leading to the final outputs [96].

nodes in the hidden layer, are called *hyperparameters*.

This simple architecture of a neural network can, given an infinite amount of diverse training data and computational resources, approximate any function to arbitrary accuracy [95]. However, in practice the performance is often improved by adding additional hidden layers, that render the architecture deeper. Complex and high dimensional data need a high number of nodes and layers, and therefore a high number of weights, to achieve good performance. Feed forward networks reach the limits of computational resources very quickly, as every weight is used for only one transition from a node to another. There are several concepts to make neural networks computationally more efficient, and by that improve their results in practice. We introduce them in the following sections.

A.2 Convolutional Neural Networks

A way to make neural networks computationally more efficient is *weight sharing*, where the same weights are reused for multiple calculations. One form of weight sharing is by using *weight kernels* of discrete convolutions. A neural network that uses this technique is called a *convolutional neural network (CNN)* [97]. This method is typically used for spatial data, where data points within each sample are strongly correlated with their neighbors, as is the case with images. The kernel $\mathcal{K}$ is a matrix, composed of weights w_{ij} that are optimized throughout the training process. To stick with the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

example of an image, the sample x is a matrix itself. The size of the kernel matrix k is smaller than the size of the sample matrix, typical kernel sizes used in computer vision are 3×3, 5×5 and 7×7. The kernel is then successively shifted across the image x, creating a *feature map* l by point-wise multiplications of the weights and the pixels of the old image:

$$l[u, v] \leftarrow \sum_{i=0}^{k} \sum_{j=0}^{k} w_{ij} x[su + i, sv + j] + B[u, v]. \tag{A.4}$$

Here s is the *stride*, a constant value that controls how many pixels the kernel is shifted. Typically s is either one, two or three. A larger stride causes a reduced dimensionality in the output image l. The matrix B is a bias matrix and also filled with weights that are optimized throughout the training process. To further reduce the dimensionality, after a convolutional layer typically a *pooling*, also called *subsampling*, layer is applied. In this layer, the average (average-pooling) or maximum (max-pooling) value over blocks of $m \times m$ pixels are taken, to form a new pixel in a feature map. As a third step after the subsampling, an activation function, typically a ReLU, is applied to each pixel. The combination of a kernel layer, a pooling layer and the activation function is called a convolutional layer. This situation is depicted in Fig. A.2. Often multiple kernels are applied in parallel, creating multiple feature maps. Also, usually multiple convolutional layers are stacked after each other, the final feature maps are then resized into one vector, which is inserted into a fully connected layer that provides the final output.

Convolutional neural networks perform best on image data, but can also be applied to time series. For complex data, CNNs with many convolutional layers, i.e. deep architecturs, work best. However, a problem intrinsic to deep architectures, is the so-called *vanishing gradient* problem [98]. Typically activation functions have a derivative smaller than one, causing problems for the backpropagation algorithm, that is responsible for the training of the network. In short, the earlier layers do not get updated, if they have many successors. There are some techniques to address this problem, two of them are applied in the ResNet, which we introduce in the following.

A.3 The ResNet

Deep neural network architectures exhibit often good capabilities to generalize, but they also suffer from the vanishing gradient problem [98], which means that the weights of early layers do not get updated. This problem is addressed by the ResNet, a very

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

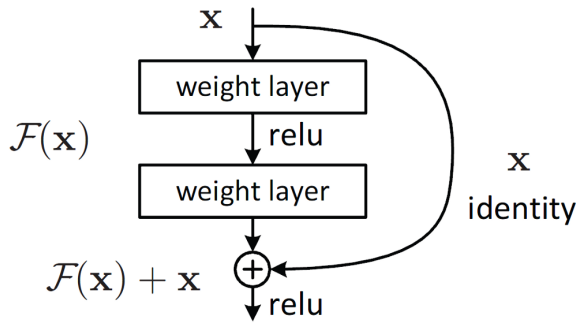


Figure A.3: The residual network uses skip connections between subsequent layers, depicted with the identity function. The output of a former layer is added to the output for a later layer. The two skipped layers are here depicted with $\mathcal{F}(\cdot)$ [99].

deep neural network architecture, that currently shows the best results for image classification [99, 100]. The ResNet uses *skip connections* between layers, to train also early layers well. This is depicted in Fig. A.3. The output x of a layer is inserted to a stack of layers $\mathcal{F}$. The output of this layer stack is then added to the input, i.e. the input is in parallel subjected to an identity function. This can be seen as a way for the network to choose how many layers it best uses to solve the approximation problem. The original ResNet architecture uses a stack of convolutional layers as weight layer stack $\mathcal{F}$. In the original paper, there are multiple architectures with different numbers of layer stacks proposed (ResNet18, ResNet34, ...).

The second method that is used by the ResNet is *batch normalization*, which also proved to achieve better optimization for early layers [101]. The batch normalization takes as input the values x within all samples over a whole mini batch $\mathcal{B} = \{x_{1...m}\}$, where m is the size of the mini batch. It first calculates the mean $\mu_\mathcal{B}$ and variance $\sigma^2_\mathcal{B}$ of all values of the samples within the mini batch:

$$\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i, \tag{A.5}$$

$$\sigma^2_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_\mathcal{B})^2 . \tag{A.6}$$

All samples x_i are then normalized, scaled and shifted:

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma^2_\mathcal{B} + \epsilon}}, \tag{A.7}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta. \tag{A.8}$$

Here ϵ is an arbitrary small constant, added for numerical stability, γ, β are two learned parameters, $\widehat{x}_i$ is the normalized sample and y_i is the final, transformed sample. In the ResNet, batch normalization is applied after every stack of convolutional layers.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

There are various adaptions and improvements of the ResNet, that even increase its performance [102]. However, an alternative to convolutions are *recurrent neural networks*. These also apply weight sharing, but with some significant differences. A special form of recurrent neural networks is explained in the following section.

A.4 Long Short-Term Memory Networks

A potential improvement to dense networks are recurrent neural networks (RNN) [103]. These take as input a time-ordered sequence $s_i = \{x_i^t\}_{t=0,\dots,T}$, which results in a splitting of the original sample into vectors x_i^t of fixed size, each corresponding to a *time step* t. The sequence is fed to the RNN consecutively. For simplicity we will only assume a network with one layer. The output of the layer is called the hidden state h. In time step t, the network's hidden state is calculated as follows:

$$h_t(x^t) = \tanh(Ux_i^t \oplus Wh_{t-1}(x^{t-1})). \tag{A.9}$$

Here U and W are weight matrices of appropriate shape. The symbol $\oplus$ denotes a concatenation of the two vectors. For every time step, the same weight matrices are applied. This can be seen as a form of time-distributed weight sharing, which is indeed the key component of RNNs. By always including the hidden state of the former time steps in the calculation, the RNN creates an internal memory. The output for later time steps in the sequence depends not only on the corresponding vector of the sequence, but also on the past ones. During the training, the backpropagation algorithm creates a gradient which, in theory, initiates learning in all former time steps. This effect is called "backpropagation in time". In practice, the backpropagation in time works well for an order of magnitude of $\mathcal{O}(5)$ time steps.

For longer backpropagation, often the vanishing gradient problem reappears. However, a special architecture of recurrent networks is capable of handling the vanishing gradient problem: The *Long Short-Term Memory (LSTM)* network [105]. Additionally to the hidden state h_t this network has an internal state, the cell state c_t. The cell state realizes even more the idea of an internal memory, being protected and read out by *gates*. The internal state of time step t in the LSTM is calculated with the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

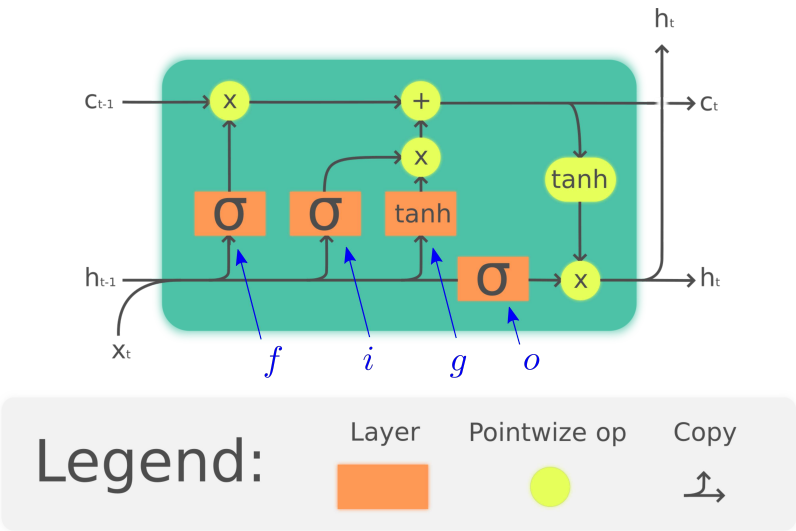


Figure A.4: Schematic of an LSTM cell. One step of the input sequence x_t is fed to the LSTM and concatenated with the hidden state of the former step h_{t-1}. The resulting state controls the input- (i), forget (f) and output (o) gate, depicted as layers (orange) with sigmoid activation function. This state is also subjected to a weight layer with tangens hyperbolicus activation function, that constructs the candidate state (g). The cell state c_t is constructed as an addition of the previous cell state c_{t-1} and the candidate state, where the input and forget gate control the share of information that is included from both. The hidden state h_t is constructed from the cell state c_t with a tangens hyperbolicus activation function and controlled by the ouput gate [104].

following equations:

$$i = \sigma\left(U^i x_t \oplus W^i h_{t-1}\right), \tag{A.10}$$

$$f = \sigma\left(U^f x_t \oplus W^f h_{t-1}\right), \tag{A.11}$$

$$o = \sigma\left(U^o x_t \oplus W^o h_{t-1}\right), \tag{A.12}$$

$$g = \tanh\left(U^g x_t \oplus W^g h_{t-1}\right), \tag{A.13}$$

$$c_t = c_{t-1} \circ f + g \circ i, \tag{A.14}$$

$$h_t = \tanh\left(c_t\right) \circ o. \tag{A.15}$$

Here $\circ$ denotes an element wise multiplication. $U^i, W^i, U^f, W^f, U^o, W^o, U^g$ and W^g denote again weight matrices of appropriate shape. σ denotes the sigmoid function. The functions i, f and o are called the input-, forget- and output-gate, respectively. They control which information from the current input is added to the cell state, which information from the past cell state is kept for the future cell state and which information from the cell state makes up the hidden state and is output as such. g is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

called the candidate state. In an RNN, g would already be the hidden state, but in an LSTM network the input gate i has yet to decide, which part of g is included to the cell state. As this mathematical framework may seem complicated at first sight, a visualization is provided in Fig. A.4.

The cell state, protected by the gates, serves as a long term memory. The gradient, produced at a certain time step, can be backpropagated for an order of magnitude of $\mathcal{O}(1000)$ time steps due to the direct connection to all former ones. This means, the LSTM network is capable of relating loss values to predictions that lie far back in the past. The LSTM architecture achieves very good results on data that consist of time series, as the data of the CRESST experiment do. It does therefore not come as a surprise, that most of our best results were achieved with LSTM neural networks (see chapter 4 and chapter 6).

We have now introduced all neural network architectures that we used for supervised learning in the work at hand. However, we also used two techniques for unsupervised learning with neural networks, which we introduce in the following sections.

A.5 Autoencoders

We already introduced autoencoders in chapter 5 and will only repeat the concept here to provide a full summary of all neural network based techniques of our work in this appendix. An autoencoder is a stack of two neural networks, an encoding network $\mathcal{E}$ and a decoding network $\mathcal{D}$ [106]. This is depicted in Fig. A.5. The encoding network maps a sample x_i to a typically lower dimensional vector l_i, the latent state:

$$\mathcal{E}(x_i) = l_i. \tag{A.16}$$

The decoding network maps the latent state back to an approximation of the original sample:

$$\mathcal{D}(l_i) \approx x_i. \tag{A.17}$$

The autoencoder, which is a stack of the decoder and encoder network, learns therefore to construct a compressed version l_i of each sample x_i, that still contains all information to reconstruct the sample. The training of an autoencoder needs no labels in the data set, it works in an unsupervised fashion. Autoencoders are used for different tasks, such as anomaly detection (see chapter 5), data compression and data simulation. For the latter, a prominent adaption, the *variational autoencoder* produces good results for many applications [107].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

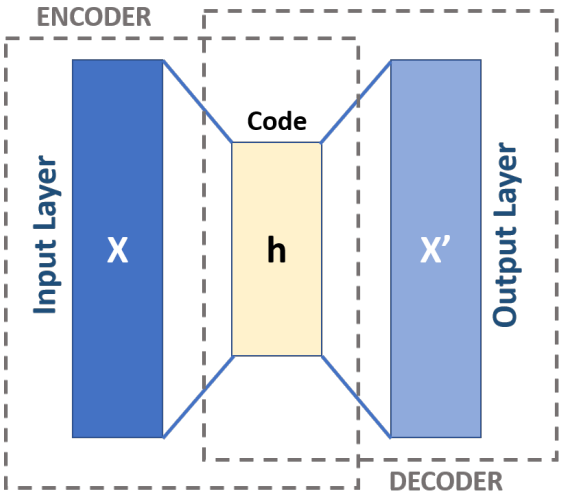


Figure A.5: The autoencoder is a stack of an encoding neural network and a decoding one. The encoder constructs a latent state h (code), which is a compressed version of the input sample x. The decoder reconstructs an approximation of the original sample x' from the latent state [108].

We shall now discuss to the last architecture, that we use in our work, that is also used for the simulation of data and was applied by us in appendix B.

A.6 Generative Adversarial Networks

Generation of artificial data, i.e. data simulation, with machine learning methods has recently gained attention, as *generative models* have shown the ability to generate realistic videos of people [109], realistic voice samples [110] and many other types of realistic fakes. The generative models that perform currently best for such applications are *generative adversarial networks (GAN)* [111]. We explain the principle of a GAN by the example of generating fake images. For that a data set of real images is needed, the task of the model is to generate fake images with similar properties.

A GAN is a stack of two neural networks, a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ (see Fig. A.6). The generator network (blue) takes as input a random noise sample and constructs a fake image from the noise. The discriminator network (red) takes alternatingly real images from the data set and fake images, generated by the generator network, as input. The real images are labeled with the number one, the fake images with the number zero. The discriminator network outputs a number between zero and one, that corresponds to its belief if the input image is real or fake. The discriminator network and the generator network are alternatingly trained. The discriminator

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

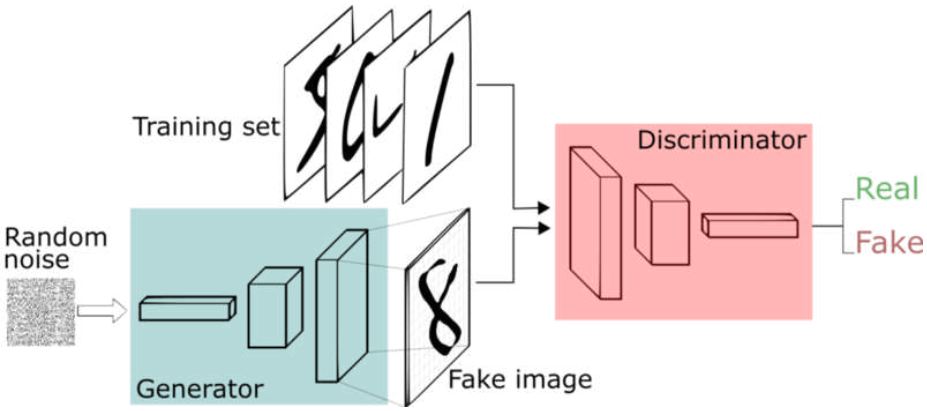


Figure A.6: The generative adversarial network consists of two opposing neural networks: The generator (blue) and the discriminator (red). The generator constructs from random noise fake samples (images in this example) that resemble the properties of a given training set. The discriminator decides which samples are real and which are fake. The networks are alternatingly trained and therefore optimized to either fool or correct each other [112].

network is trained as a binary classifier, to correctly discriminate between real and fake images. The generator is trained to generate images that the discriminator believes to be real, i.e. that cause values close to one on the output of the discriminator. By that, both networks are optimized over time. This situation is called a Minimax (zero-sum, non-cooperative) game in game theory. In the ideal case, the opposing networks $\mathcal{G}, \mathcal{D}$ reach a Nash equilibrium, an ideal point $\mathcal{G}^*$ in the objective function $v(\mathcal{G}, \mathcal{D})$, i.e.

$$\mathcal{G}^* = \arg\min_{\mathcal{G}} \max_{\mathcal{D}} v(\mathcal{G}, \mathcal{D}). \tag{A.18}$$

Here the ideal point $\mathcal{G}^*$ corresponds to a generator, that produces fake images that are indistinguishable from real images. The default choice of the objective function is

$$v\left(\boldsymbol{\theta}^{(\mathcal{G})}, \boldsymbol{\theta}^{(\mathcal{G})}\right) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\log \mathcal{D}(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{model}}}\left[\log(1-\mathcal{D}(\boldsymbol{x}))\right], \tag{A.19}$$

where $\boldsymbol{\theta}^{(\mathcal{G})}$ corresponds to the weights of the generator network, $\boldsymbol{\theta}^{(\mathcal{D})}$ to those of the discriminator network, p_{data} is the distribution of the real images in the data set, p_{model} is the distribution of the fake images from the generator network and $\mathcal{D}(\boldsymbol{x})$ is the output between zero and one of the discriminator network, for a given input $\boldsymbol{x}$.

We can observe a similarity between the architectures of an autoencoder and the GAN. Both models consist of two networks, the architecture of a generator network is often similar to that of a decoder network, the architecture of a discriminator is often similar to that of an encoder. As mentioned above, an adaption of the autoencoder, the variational autoencoder, is also used for the generation of fake data. Autoencoder

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

models are typically easier to train and more robust, while GANs typically exhibit more realistic results, but are harder to train.

We gave in this appendix a condensed introduction to deep learning and the neural network models we use in our work. For more detailed explanations we recommend Ref. [69] and [70]. The two currently most widely used frameworks are Keras, based on Tensorflow, and PyTorch. For our work, we used the PyTorch framework.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Appendix B

Simulated Noise Baselines with Generative Adversarial Networks

We introduced a method for the simulation of noise baselines in chapter 3. In this appendix we study the results of the simulation of noise baselines with generative adversarial networks (GANs), which we introduced in appendix A. Our requirements for the simulated baselines are

- the same noise power spectrum as in the measured baselines, and
- the same correlated structure as in the measured baselines, especially regarding the drift structure of the baselines.

A sample of eight measured baselines from the TUM40 phonon detector is shown in Fig. B.1. We can see mostly linear and quadratic drifts, although a clear statement is difficult due to the overlaid noise.

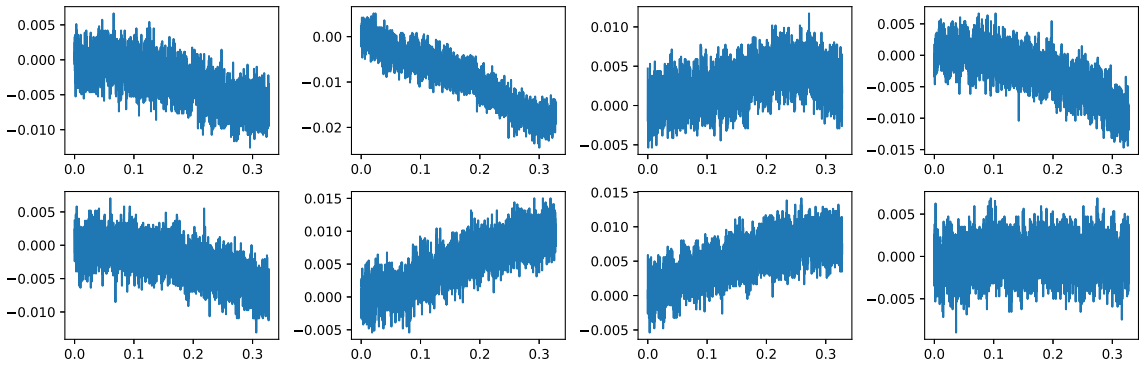


Figure B.1: Measured noise baselines of the TUM40 phonon detector, used in run 33. All baselines are plotted such that their first value equals zero. The values on the abscissa (x-axis) have dimension s, the values on the ordinate (y-axis) have dimension V.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

For the simulation with GANs, we construct a data set of 5000 measured baselines, clear them with the RMS fit from all accidentally recorded artifacts, as we did in chapter 3, and train a GAN to generate similar baselines. A sample of 25 generated baselines is shown in Fig. B.2. We observe drift structures similar to those of the measured baselines, however, we also observe a peculiar frequency of roughly 17 small spikes in almost all baselines.

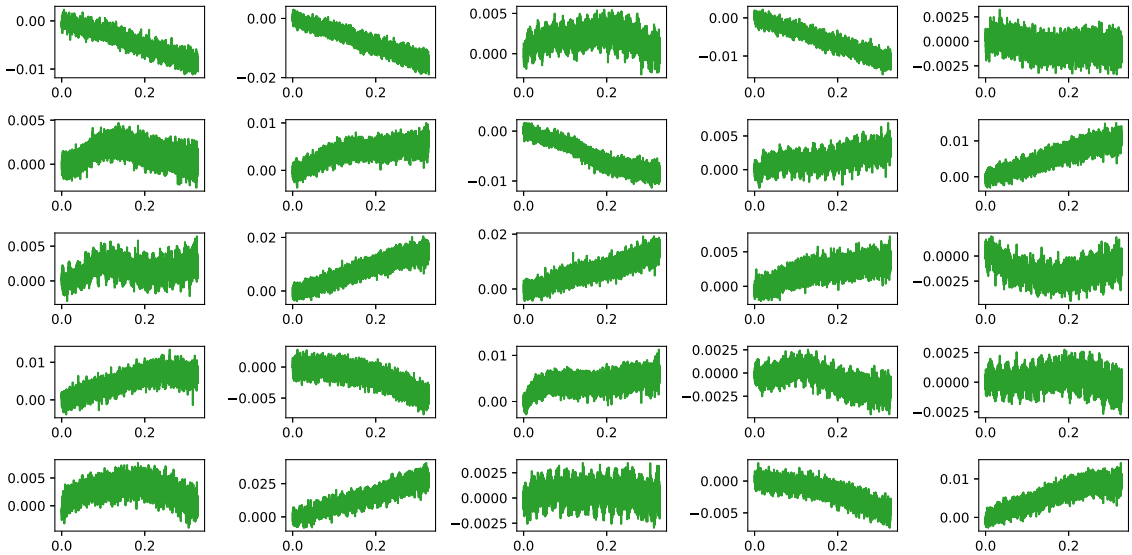


Figure B.2: Generated noise baselines with a GAN, modeled after the measured noise baselines from the TUM40 detector (Fig. B.1). The values on the abscissa (x-axis) have dimension s, the values on the ordinate (y-axis) have dimension V.

If we compare the noise power spectra of the measured baselines (Fig. B.3, blue) and the GAN-generated baselines (Fig. B.3, orange), we can see that the generated baselines mimic all the characteristic spikes that occur in the spectrum, but have overall a lower amplitude in all other frequencies. However, the characteristic peaks around 400 Hz, 500 Hz and 800 Hz only slightly visible in the simulated spectrum. The peak at 50 Hz is even higher than in the original baselines. This explains also the small spikes, that we observe in the generated baselines: The frequency of 17 spikes per 330 ms baseline corresponds exactly to the 50 Hz peak, that we observe in the noise power spectrum. These spikes also exist in the measured baselines, but are not visible due to the overlay with other noise, of other frequencies. It should not come as a surprise, that the GAN lowers the amplitude of stochastic noise. Neural network are approximation techniques and all approximation techniques do, to some extend, smear out noise fluctuations. Usually this reduction of noise fluctuations is exactly

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

the desired property for which one uses approximation techniques, however, as we will see shortly, in our case it perishes the noise power spectra of the simulated baselines.

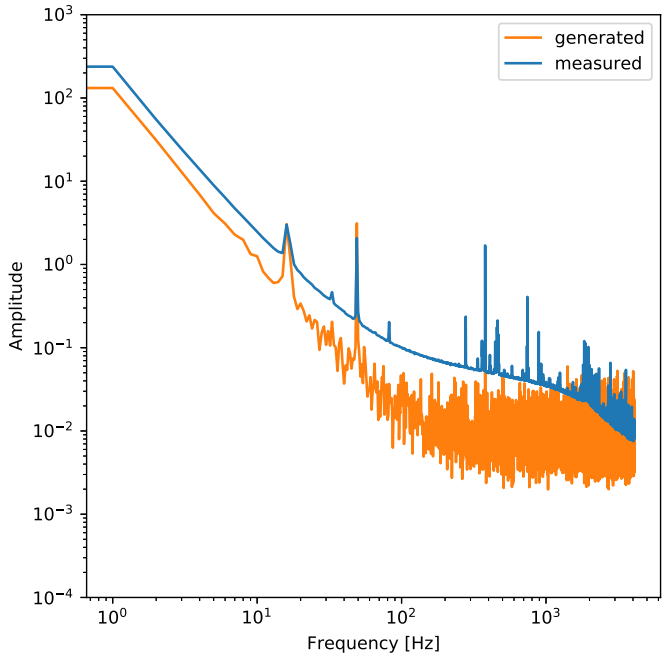


Figure B.3: Noise power spectrum of the GAN generated baselines (orange) and the measured baselines (blue). We see an overall lower amplitude in the generated spectrum, while it characteristic 50 Hz-peak has an even higher amplitude than in the measured spectrum. The peak around 20 Hz is very well reconstructed. The high-frequent fluctuations above 400 Hz are only slightly visible in the generated spectrum.

Even though the GAN-generated noise baselines mimic all the characteristic features of the measured baselines, they do not have the same noise power spectrum, which is an important property for some of the methods used in CRESST analysis. For our work, we therefore prefer to use the non-machine learning based simulation technique that we introduced in chapter 3. These baselines definitely have the same noise power spectrum as the measured ones, since they were superposed with noise such that the spectra match. For comparison, some of those simulated baselines are shown in Fig. B.4. However, for future simulations GANs could still prove useful, e.g. for the quick generation of more data, for which no knowledge-based model as our current model for the baseline simulation is available. Also, our baseline model does ignore possible correlations in the phases of frequencies, which could still be a shortcoming of the simulation. The GAN simulation could be improved with a subsequent noise-overlay, to account for the differences in the noise power spectrum,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

as it was done in our baseline simulation with the drift structure. For a more detailed explanation of the GAN-based simulation method see Ref. [113].

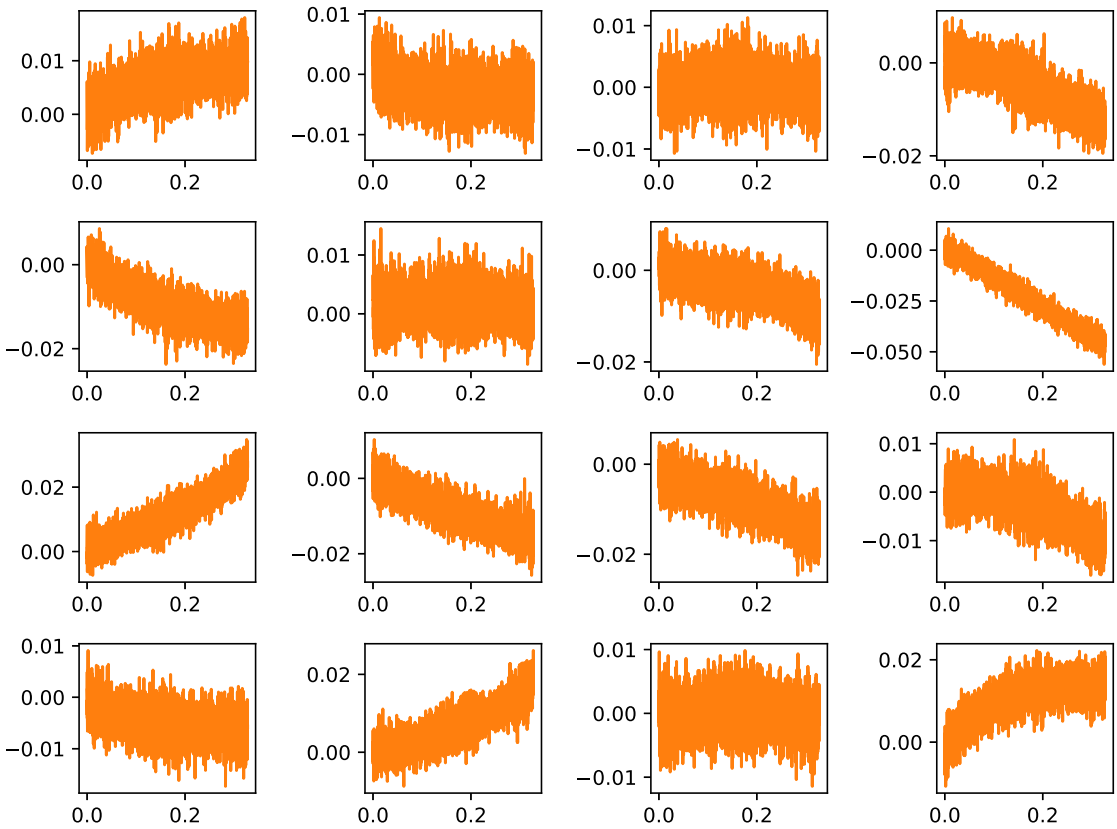


Figure B.4: Simulated noise baselines with the baseline simulation, introduced in chapter 3. The values on the abscissa (x-axis) have dimension s, the values on the ordinate (y-axis) have dimension V.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

References

[1] F. Zwicky. Republication of: The redshift of extragalactic nebulae. *General Relativity and Gravitation*, 41(1):207–224, January 2009.

[2] Vera C. Rubin and W. Kent Ford, Jr. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *The Astrophysical Journal*, 159:379, February 1970.

[3] Planck Collaboration. Planck 2013 results. XVI. Cosmological parameters. *Astronomy & Astrophysics*, 571:A16, November 2014. arXiv: 1303.5076.

[4] J. Jochuma, G. Angloher, C. Bucci, S. Cooper, C. Cozzini, P. DiStefano, F. von Feilitzsch, T. Frank, D. Hauff, Th. Jagemann, R. Keeling, H. Kraus, M. Loidl, O. Meier, U. Nagel, F. Pröbst, Y. Ramachers, A. Rulofs, J. Schnagl, W. Seidel, I. Sergeyev, M. Sisti, M. Stark, S. Uchaikin, L. Stodolsky, and H. Wulandari. Limits on WIMP dark matter using sapphire cryogenic detectors. *Nuclear Physics B - Proceedings Supplements*, 124:189–192, July 2003.

[5] G. Angloher, P. Carniti, L. Cassina, L. Gironi, C. Gotti, A. Gütlein, D. Hauff, M. Maino, S. S. Nagorny, L. Pagnanini, G. Pessina, F. Petricca, S. Pirro, F. Pröbst, F. Reindl, K. Schäffner, J. Schieck, and W. Seidel. The COSINUS project: perspectives of a NaI scintillating calorimeter for dark matter search. *The European Physical Journal C*, 76(8):441, August 2016.

[6] The NUCLEUS Collaboration. NUCLEUS: Exploring Coherent Neutrino-Nucleus Scattering with Cryogenic Detectors. *Journal of Low Temperature Physics*, 199(1-2):433–440, April 2020.

[7] F. Pröbst, M. Frank, S. Cooper, P. Colling, D. Dummer, P. Ferger, G. Forster, A. Nucciotti, W. Seidel, and L. Stodolsky. Model for cryogenic particle detectors with superconducting phase transition thermometers. *Journal of Low Temperature Physics*, 100(1-2):69–104, July 1995.

[8] CRESST Collaboration. First results from the CRESST-III low-mass dark matter program. *Physical Review D*, 100(10):102002, November 2019. arXiv: 1904.00498.

[9] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015.

[10] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8:42200–42216, 2020.

[11] Janis Fluri, Tomasz Kacprzak, Aurelien Lucchi, Alexandre Refregier, Adam Amara, Thomas Hofmann, and Aurel Schneider. Cosmological constraints with deep learning from KiDS-450 weak lensing maps. *Physical Review D*, 100(6):063514, September 2019.

[12] Charanjit K. Khosa, Lucy Mars, Joel Richards, and Veronica Sanz. Convolutional Neural Networks for Direct Detection of Dark Matter. *arXiv:1911.09210 [astro-ph, physics:hep-ph]*, November 2019. arXiv: 1911.09210.

[13] Christoph Dreissigacker, Rahul Sharma, Chris Messenger, Ruining Zhao, and Reinhard Prix. Deep-learning continuous gravitational waves. *Physical Review D*, 100(4):044009, August 2019.

[14] Christoph Mühlmann. Pulse-shape discrimination with deep learning in cresst. Unpublished Master Thesis, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[15] David Matravers. Steven Weinberg: Cosmology: Oxford University Press, 2008, 544p., GBP45.00, ISBN13: 978-0-19-852682-7. *General Relativity and Gravitation*, 41(6):1455–1458, June 2009.

[16] Peter J. Bussey. Introduction to cosmology, 2nd edition, by b. ryden. *Contemporary Physics*, 58(4):381–381, 2017.

[17] Albert Einstein. The Field Equations of Gravitation. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)*, 1915:844–847, 1915.

[18] Jochen Schieck and Holger Kluck. Dark matter. Unpublished Lecture Notes, 2019.

[19] F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. *The Astrophysical Journal*, 86:217, October 1937.

[20] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 142:419, July 1965.

[21] TImm Wrase. Cosmology and particle physics. Unpublished Lecture Notes, 2019.

[22] European Space Agency. Planck's view of the cosmic microwave background, 2018. [Online; accessed 21-April-2020].

[23] Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Jörg Colberg, and Frazer Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, June 2005.

[24] Douglas Clowe, Maruša Bradač, Anthony H. Gonzalez, Maxim Markevitch, Scott W. Randall, Christine Jones, and Dennis Zaritsky. A Direct Empirical Proof of the Existence of Dark Matter. *The Astrophysical Journal*, 648(2):L109–L113, September 2006.

[25] E. Corbelli and P. Salucci. The extended rotation curve and the dark matter halo of M33. *Monthly Notices of the Royal Astronomical Society*, 311(2):441–447, January 2000.

[26] Wikipedia, The Free Encyclopedia. Galaxy rotation curve. [Online; accessed 19-April-2020].

[27] Kim Griest. The Search for the Dark Matter: WIMPs and MACHOs. *Annals of the New York Academy of Sciences*, 688(1):390–407, June 1993.

[28] Marco Drewes. THE PHENOMENOLOGY OF RIGHT HANDED NEUTRINOS. *International Journal of Modern Physics E*, 22(08):1330019, August 2013.

[29] Athanasios Lahanas. LSP as a Candidate for Dark Matter. In Lefteris Papantonopoulos, editor, *The Invisible Universe: Dark Matter and Dark Energy*, volume 720, pages 35–68. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. Series Title: Lecture Notes in Physics.

[30] M. Milgrom. A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *The Astrophysical Journal*, 270:365, July 1983.

[31] Florian Reindl. Exploring light dark matter with cresst-ii low-threshold detectors. Unpublished PhD Thesis, 2016.

[32] XENON Collaboration. Search for Light Dark Matter Interactions Enhanced by the Migdal Effect or Bremsstrahlung in XENON1T. *Physical Review Letters*, 123(24):241803, December 2019.

[33] M. Klasen, M. Pohl, and G. Sigl. Indirect and direct search for dark matter. *Progress in Particle and Nuclear Physics*, 85:1–32, November 2015.

[34] Gianfranco Bertone and Dan Hooper. History of dark matter. *Reviews of Modern Physics*, 90(4):045002, October 2018.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[35] Rafael F Lang and Wolfgang Seidel. Search for dark matter with CRESST. *New Journal of Physics*, 11(10):105017, October 2009.

[36] G. Angloher, A. Bento, C. Bucci, L. Canonica, A. Erb, F. von Feilitzsch, N. Ferreiro Iachellini, P. Gorla, A. Gütlein, D. Hauff, P. Huff, J. Jochum, M. Kiefer, C. Kister, H. Kluck, H. Kraus, J.-C. Lanfranchi, J. Loebell, A. Münster, F. Petricca, W. Potzel, F. Pröbst, F. Reindl, S. Roth, K. Rottler, C. Sailer, K. Schäffner, J. Schieck, J. Schmaler, S. Scholl, S. Schönert, W. Seidel, M. von Sivers, L. Stodolsky, C. Strandhagen, R. Strauss, A. Tanzke, M. Uffinger, A. Ulrich, I. Usherov, S. Wawoczny, M. Willers, M. Wüstrich, and A. Zöller. Results on low mass WIMPs using an upgraded CRESST-II detector. *The European Physical Journal C*, 74(12):3184, December 2014.

[37] Jens Michael Schmaler. *The CRESST Dark Matter Search - New Analysis Methods and Recent Results*. Phd thesis, TU München, 2010.

[38] R. Strauss, G. Angloher, A. Bento, C. Bucci, L. Canonica, A. Erb, F. von Feilitzsch, N. Ferreiro, P. Gorla, A. Gütlein, D. Hauff, J. Jochum, M. Kiefer, H. Kluck, H. Kraus, J.-C. Lanfranchi, J. Loebell, A. Münster, F. Petricca, W. Potzel, F. Pröbst, F. Reindl, S. Roth, K. Rottler, C. Sailer, K. Schäffner, J. Schieck, S. Scholl, S. Schönert, W. Seidel, M. von Sivers, M. Stanger, L. Stodolsky, C. Strandhagen, A. Tanzke, M. Uffinger, A. Ulrich, I. Usherov, S. Wawoczny, M. Willers, M. Wüstrich, and A. Zöller. A detector module with highly efficient surface-alpha event rejection operated in CRESST-II Phase 2. *The European Physical Journal C*, 75(8):352, August 2015.

[39] Nahuel Ferreiro Iachellini. Increasing the sensitivity to low mass dark matter in cresst-iii with a new daq and signal processing. Unpublished PhD Thesis, 2018.

[40] S Di Domizio, F Orio, and M Vignati. Lowering the energy threshold of large-mass bolometric detectors. *Journal of Instrumentation*, 6(02):P02007–P02007, February 2011.

[41] Martin Stahlberg. Increasing the sensitivity to low mass dark matter in cresst-iii with a new daq and signal processing. Unpublished PhD Thesis, 2020.

[42] S. Yellin. Finding an upper limit in the presence of an unknown background. *Physical Review D*, 66(3):032005, August 2002.

[43] Daniel Schmiedmayer. Calculation of dark-matter exclusion-limits using a maximum likelihood approach. Unpublished Master Thesis, 2019.

[44] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip Reading Sentences in the Wild. *arXiv:1611.05358 [cs]*, January 2017. arXiv: 1611.05358.

[45] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016. arXiv: 1609.03499.

[46] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*, October 2016. arXiv: 1609.08144.

[47] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):1–13, July 2017.

[48] David Ha and Douglas Eck. A Neural Representation of Sketch Drawings. *arXiv:1704.03477 [cs, stat]*, May 2017. arXiv: 1704.03477.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[49] Zbigniew Wojna, Alex Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based Extraction of Structured Information from Street View Imagery. *arXiv:1704.03549 [cs]*, August 2017. arXiv: 1704.03549.

[50] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv:1706.05125 [cs]*, June 2017. arXiv: 1706.05125.

[51] European Space Agency. Planck's view of the cosmic microwave background, 2018. [Online; accessed 21-April-2020].

[52] P. R. Strickland. The thermal equivalent circuit of a transistor. *IBM Journal of Research and Development*, 3(1):35–45, 1959.

[53] M. Carrettoni and O. Cremonesi. Generation of noise time series with arbitrary power spectrum. *Computer Physics Communications*, 181(12):1982–1985, December 2010.

[54] Johannes Rothe. Achieving low thresholds: Cryogenic detectors for low-mass dark matter searches. Unpublished Master Thesis, 2016.

[55] Felix Wagner. Magnetization of superconducting atomchips. Unpublished Bachelor Thesis, 2018.

[56] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, September 1956.

[57] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.

[58] Alan V. Oppenheim, Alan S. Willsky, and Ian T. Young. *Signals and systems*. Prentice-Hall signal processing series. Prentice-Hall, Englewood Cliffs, N.J, 1983.

[59] S. Riboldi, R. Abbiati, A. Geraci, and E. Gatti. Experimental comparison of state-of-the-art methods for digital optimum filter synthesis with arbitrary constraints and noise. *IEEE Transactions on Nuclear Science*, 52(4):954–958, August 2005.

[60] J. Dongarra and F. Sullivan. Guest Editors Introduction to the top 10 algorithms. *Computing in Science & Engineering*, 2(1):22–23, January 2000.

[61] Peter Frolkovič. Numerical recipes: The art of scientific computing: William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling: Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, Sydney, 1986, 20+818 pp. *Acta Applicandae Mathematicae*, 19(3):297–299, June 1990.

[62] Wikipedia, The Free Encyclopedia. Matched filter. [Online; accessed 19-April-2020].

[63] Rudolf Frühwirth. Wahrscheinlichkeitsrechung und statistik für studierende der physik. Unpublished Lecture Notes.

[64] Stéphane Lathuilière, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. A Comprehensive Analysis of Deep Regression. *arXiv:1803.08450 [cs]*, February 2019. arXiv: 1803.08450.

[65] Hunter Gabbard, Michael Williams, Fergus Hayes, and Chris Messenger. Matching Matched Filtering with Deep Networks for Gravitational-Wave Astronomy. *Physical Review Letters*, 120(14):141103, April 2018.

[66] Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, and Ramon Winterhalder. How to GAN away Detector Effects. *arXiv:1912.00477 [hep-ph]*, March 2020. arXiv: 1912.00477.

[67] Philipp Neuhuber. Advancing the smodels framework with neural networks. Unpublished Master Thesis, 2020.

[68] Felix Wagner. Approximation of cross section upper limit maps with neural networks. Unpublished Bachelor Thesis, 2018.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[69] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016.

[70] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, 810:1–124, May 2019.

[71] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. arXiv: 1412.6980.

[72] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989.

[73] Tong Yu and Hong Zhu. Hyper-Parameter Optimization: A Review of Algorithms and Applications. *arXiv:2003.05689 [cs, stat]*, March 2020. arXiv: 2003.05689.

[74] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, April 2020.

[75] Karl Pearson. LIII. *On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901.

[76] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

[77] Ian Jolliffe. Principal Component Analysis. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[78] Rodolphe Priam. Symmetric Generative Methods and tSNE: A Short Survey:. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 356–363, Funchal, Madeira, Portugal, 2018. SCITEPRESS - Science and Technology Publications.

[79] David M. Chan, Roshan Rao, Forrest Huang, and John F. Canny. t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data. *arXiv:1807.11824 [cs, stat]*, July 2018. arXiv: 1807.11824.

[80] Guijuan Zhang, Yang Liu, and Xiaoning Jin. A survey of autoencoder-based recommender systems. *Frontiers of Computer Science*, 14(2):430–450, April 2020.

[81] Mahbubul Alam, Manar D. Samad, Lasitha Vidyaratne, Alexander Glandon, and Khan M. Iftekharuddin. Survey on Deep Neural Networks in Speech and Vision Systems. *arXiv:1908.07656 [cs, eess, stat]*, November 2019. arXiv: 1908.07656.

[82] Malik Khizar Hayat, Ali Daud, Abdulrahman A. Alshdadi, Ameen Banjar, Rabeeh Ayaz Abbasi, Yukun Bao, and Hussain Dawood. Towards Deep Learning Prospects: Insights for Social Media Analytics. *IEEE Access*, 7:36958–36979, 2019.

[83] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, April 2020.

[84] Katarzyna Janocha and Wojciech Marian Czarnecki. On Loss Functions for Deep Neural Networks in Classification. *arXiv:1702.05659 [cs]*, February 2017. arXiv: 1702.05659.

[85] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]*, October 2016. arXiv: 1506.02142.

[86] Christopher M. Bishop. *Bayesian methods for neural networks*. Oxford University Press, United Kingdom, 1995.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[87] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv:2001.07685 [cs, stat]*, January 2020. arXiv: 2001.07685.

[88] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018.

[89] Sarah L. Harris and David Money Harris. Sequential Logic Design. In *Digital Design and Computer Architecture*, pages 108–171. Elsevier, 2016.

[90] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.

[91] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[92] Wikipedia, The Free Encyclopedia. Artificial neural network. [Online; accessed 19-April-2020].

[93] Herbert E. Robbins. A stochastic approximation method. 2007.

[94] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[95] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, December 1989.

[96] Wikipedia, The Free Encyclopedia. Convolutional neural network. [Online; accessed 19-April-2020].

[97] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.

[98] Josef Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. Unpublished Master Thesis, 1991.

[99] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

[100] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.

[101] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, March 2015. arXiv: 1502.03167.

[102] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv:2004.08955 [cs]*, April 2020. arXiv: 2004.08955.

[103] Michael I. Jordan. Artificial neural networks. chapter Attractor Dynamics and Parallelism in a Connectionist Sequential Machine, pages 112–127. IEEE Press, Piscataway, NJ, USA, 1990.

[104] Wikipedia, The Free Encyclopedia. Long short-term memory. [Online; accessed 19-April-2020].

[105] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[106] Dana H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1*, AAAI'87, page 279–284. AAAI Press, 1987.

[107] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114.

[108] Wikipedia, The Free Encyclopedia. Autoencoder. [Online; accessed 19-April-2020].

[109] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *arXiv:1905.08233 [cs]*, September 2019. arXiv: 1905.08233.

[110] Yang Gao, Rita Singh, and Bhiksha Raj. Voice Impersonation using Generative Adversarial Networks. *arXiv:1802.06840 [cs, eess]*, February 2018. arXiv: 1802.06840.

[111] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[112] Thalles Silva. An intuitive introduction to generative adversarial networks (gans), 2020. [Online; accessed 19-April-2020].

[113] Pascal Schanner. Discriminating between measured and simulated baselines. Unpublished Bachelor Thesis, 2020.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Index

$CaWO_4$, 18

Absorber crystal, 69
Accuracy, 87
Activation function, 56, 100
Autoencoder, 80, 107

Backpropagation, 57, 101
Baryonic matter, 5
Batch normalization, 104
Big Bang nucleosynthesis, 7
Black hole, 9
Bolometer, 31
Bullet cluster, 8

Calorimeter, 31
Carousel, 16
Carrier crystal, 18, 69
Carrier recoil event, 23, 69, 86
Classification problem, 28, 100
Coincidence cut, 23
Cold dark matter, 9
Cold finger, 16
Coma cluster, 5
Composite design, 69
Control pulses, 22
Convolutional neural network (CNN), 59, 103
Cosmic microwave background (CMB), 5
Cut, 23

Dark energy, 5
Dark matter, 5
Dark matter halo, 9
Deep learning, 28, 56, 99
Dense neural network, 100
Detector A, 51
Detector module, 17
Direct detection, 11

Efficiencies, 24
Electrothermal feedback, 41
Energy calibration, 22
Energy-voltage conversion (EPC) rate, 22
Energy-voltage conversion rate, 90
Energy-voltage conversion rate (EPC), 51
Ensemble learning, 62

Epoch, 57, 101
Exclusion limit, 13

Feed forward neural network (FNN), 100
Fully connected neural network, 100
Function approximator, 56, 100

Gate, 106
General quality cut, 68
Generative adversarial network (GAN), 108, 111
Generative models, 107, 108
Gradient descent, 57, 101
Gravitational lensing, 8

Heater pulses, 40
Hidden state, 57, 101
Hyperparameter, 101
Hyperparameters, 58

Inflation, 7

Kernel, 103
Knowledge-based methods, 27
Kullback-Leibler divergence, 75

Label, 27, 100
Latent state, 57, 107
Layer, 57, 101
LCDM-model, 5
Learning rate, 57, 101
Light detector, 18
Light-yield parameter, 20
Light-yield plot, 26
Lightest supersymmetric particle (LSP), 10
Long Short-Term Memory (LSTM) neural network, 59, 106
Loss function, 57, 101
LSTM carrier cut, 87
LSTM noise cut, 90

Machine learning, 27
MACHOs (massive compact halo object), 9
Massive sterile neutrinos, 10
Michael light detector, 18
Mini batch, 57, 101
MOND, 10

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Moving average, 46
Muon veto, 17

Neural network, 56, 99
Neutrinos, 9
Neutrons, 17
Node, 101
Noise baseline, 32
Noise Baselines, 111
Noise power spectrum (NPS), 33
Non-thermal component, 30

One-hot encoding, 100
Operating point, 22
Optimum filter, 21, 46
Overfitting, 58, 100

Perplexity, 74
Phonon detector, 18
Pooling, 103
Pre-Trigger, 32
Prediction, 100
Principal component analysis (PCA), 72
Pulse shape model, 30

Quality cut, 23, 68

Radon box, 17
Rate cut, 23
Rectified linear unit (ReLU), 57, 100
Recurrent neural network (RNN), 105
Regression problem, 28, 100
Reinforcement learning, 99
ResNet, 59, 104
Rotation curves, 9
Run, 22

Sample, 27, 28
Sample frequency, 21
Saturated Event, 38
Scattering rate, 11
Self-heating, 41
Shannon entropy, 74
Shielding, 17
Silicon-on-sapphire (SOS) disk, 18
Simulated data set, 37
Skip Connections, 104
Specific quality cut, 68
Stability cut, 23
Standard event, 21, 31
Standard event fit, 21, 51, 78
Standard model of cosmology, 5
Stride, 103
Structure formation, 7
Subsampling, 103
Superconducting Quantum Interference Device (SQUID), 19
Supersymmetry (SUSY), 10
Supervised learning, 27, 99

t-distributed stochastic neighbor embedding (t-SNE), 74
Test pulses, 22, 32, 40
Test set, 58, 100
Thermal component, 30
Training, 57, 100
Training set, 58, 100
Transition curve, 19
Transition edge sensor (TES), 18
Trigger-walk, 37
Truncated fit, 39
TUM 40 phonon detector, 51
TUM40 phonon detector, 18

Unsupervised learning, 27, 99

Validation set, 58, 100
Vanilla neural network, 100
Vanishing gradient problem, 104, 107
Variational autoencoder, 107
Velocity distribution of galaxies, 5

Weight sharing, 103
Weights, 57, 100
WIMP (weakly interactive massive particle), 10

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Acknowledgements

I want to thank HEPHY and the Austrian academy of sciences for the provided ressources, particularly the computing power, which was essential for my work. Especially I want to thank Jochen, my supervisor, our group leader and the director of our institute. Due to his engagement in the past years, he established a Rare Event Search group at HEPHY and I had the opportunity to join it for my master thesis. Also, it should not be taken as granted, that a master student may travel to the experiment and attend collaboration meetings and summer schools. I am grateful for the trust he put in me.

Truly from the bottom of my heart I want to thank Wolfgang and Florian, who shared my content-related supervision in the fields of machine learning and dark matter physics. I cannot wish for more mentoring, encouraging and patient supervisors and I am excited to work with them for three more years.

A substantial part of my learning experience in the past year was accompanied by my fellow members of our research group. I am very thankful for the productive and merry atmosphere, in which I felt welcome from the first day on.

I want to thank my parents for the year-long personal support and for being role models in so many things. I also want to thank my older brother, who taught me the most about the things that really matter in life and science, and continues to do so.

For the biggest contribution to my academic success I want to thank my study group. Spending every free minute during the semester at the university, calculating homework exercises, is far more bearable when it is done together with good friends.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub