Karlsruhe Institute of Technology

ETP-KA/2025-11

Search for Axion-Like Particles in $e^+e^- \to e^+e^-\gamma\gamma$ Events at Belle II

Frederik Schmitt

Master Thesis

at the Department of Physics
Institute of Experimental Particle Physics (ETP)

Supervisor:	Prof. Dr. Torben Ferber
Co-Supervisor:	Prof. Dr. Markus Klute
Advisor:	M.Sc. Alexander Heidelbach
Second advisor:	Dr. Giacomo De Pietro

01.10. 2024 – 30.09. 2025

Karlsruher Institut für Technologie
Fakultät für Physik
D-76128 Karlsruhe

Karlsruher Institut für Technologie

ETP-KA/2025-11

Suche nach axionartigen Teilchen in $e^+e^- \longrightarrow e^+e^-\gamma\gamma$ Ereignissen bei Belle II

Frederik Schmitt

Masterthesis

an der Fakultät für Physik
Institut für Experimentelle Teilchenphysik (ETP)

Referent:	Prof. Dr. Torben Ferber
Korreferent:	Prof. Dr. Markus Klute
Betreuender Mitarbeiter:	M.Sc. Alexander Heidelbach
Zweiter betreuender Mitarbeiter:	Dr. Giacomo De Pietro

01.10. 2024 – 30.09. 2025

Karlsruher Institut für Technologie
Fakultät für Physik
D-76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.
Karlsruhe, 30.09.2025

. . . . . .
(Frederik Schmitt)

Disclaimer

Collaborative Usage of Resources

As modern particle physics requires complex setups, both experimentally and in computing, parts of this thesis rely on previous work done by the Belle II collaboration, and the Belle II software computing and data production groups.
The SuperKEKB particle accelerator, which provides the beams for the Belle II experiment, was built and is operated by the SuperKEKB accelerator group. The Belle II detector was designed and is operated by the Belle II collaboration. The collaborative framework Belle II Analysis Software Framework (basf2), developed and maintained by the Belle II collaboration, is introduced and credited in section 3.1.
The data production group simulates background events centrally and distributes them collaboratively as credited in section 3.3. These samples are used by me and not self-generated. Signal is privately generated using the event generator MadGraph[1].

The original idea, implementation, and structure of the analysis code pipeline, including the plotting and fitting framework, were developed by Alexander Heidelbach for his PhD thesis. In close collaboration, we extended this pipeline to create a toolset that can be applied more generally to analyses with a similar structure. While attributing individual contributions in detail is not straightforward, the analysis code's Git [2] repository [3] provides a transparent record of authorship and development. All results, plots, fits, and their interpretation presented in this thesis were produced as part of my own work.

Usage of Generative Artificial Intelligence

This thesis uses Grammarly[1], a tool for language correction, spell checking, and rephrasing. Beyond exchanging words or checking typing, the text for this thesis is entirely written by me.
Additionally, ChatGPT[2], a generative language model, is used for code work. This is limited to minor tasks, such as introductions to new Python modules like Polars, code refactoring, or fixing minor issues like type inconsistencies. Whenever any code from it is used, it is thoroughly tested and validated to ensure it performs as expected.

[1] `https://app.grammarly.com/`, Accessed: 2025.09.23
[2] `https://chatgpt.com/`, Accessed: 2025.09.23

Contents

1. Introduction

In 1925, the mathematical formulation of Quantum Mechanics (QM) revolutionized our scientific understanding of the microscopic world. A century later, its legacy continues to shape modern particle physics. While QM successfully described matter at the smallest scales, it lacked predictive power to explain the fundamental interactions underlying high-energy processes. This transition, from the quantum description of particles to the theory of their interactions, was achieved half a century later with the development of the Standard Model (SM), a unified and predictive theory of particle physics. Probably the most famous example of this predictive power is the discovery of the Higgs Boson in 2012 [4, 5], which was already theoretically predicted in 1964 [6]. Today, the SM predictions are highly accurate in many areas of particle physics across multiple orders of energy, and all attempts to find deviations have shown none or very little effects. However, we know it is incomplete, as some effects are not included or not adequately explained. To resolve the shortcomings of the SM theorists proposed numerous Beyond the Standard Model (BSM) theories, many including new particles.

Axion-Like Particles (ALPs), as analyzed in this thesis, present compelling BSM scenarios with a yet mostly unexplored range in phase space, particularly the GeV/c^2 ALP-mass range. Due to the high collision rate and the high precision detector, the Belle II experiment is an ideal environment to search for GeV/c^2 ALPs at a particle collider experiment. While the ALP can decay into multiple signatures, this work focuses on a model where it decays predominantly into two resolved photons. Experimentally, this allows the search for a peak in the invariant diphoton mass $M_{\gamma\gamma}$ from an ALP decay signal over the SM background. Belle II already published a search for the same ALP model; however, this is the first effort to perform an experimental sensitivity study for the t-channel $\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma$ process. The search for the t-channel process is motivated by two contrasting phenomenology papers, one stating that the analysis is not feasible [7], and one that it yields comparable or better sensitivity than the previously used s-channel process $\mathrm{e}^+\,\mathrm{e}^- \to \gamma\,a,\ a \to \gamma\gamma$ [8]. In this work, I present an analysis strategy that demonstrates the feasibility of analyzing the fully reconstructed $\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma$ process at Belle II and yields comparable sensitivity to that obtained from analyzing $\mathrm{e}^+\,\mathrm{e}^- \to \gamma\,a,\ a \to \gamma\gamma$ events.

After a brief introduction to the theory of ALPs, especially those with photon couplings and the present experimental results, in chapter 2, chapter 3 explains the Belle II experiment and the collaborative basf2 used. Providing a general overview of the proposed analysis strategy is the topic in chapter 4, and the subsequent chapters elaborate on the individual steps of this analysis workflow in greater detail. Chapter 5 discusses the dataset used for Monte Carlo (MC) events, which are privately generated and reconstructed, as well as

the centrally available SM background. The reconstruction of events and used selections are explained in chapter 6, which also includes discussions on the dominant background processes, resulting background spectrum, and signal efficiencies. A Neural Network (NN) event classifier for further suppression of SM background events is explained in chapter 7. With most of the SM events removed, the expected sensitivity on the cross section of the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process is derived in chapter 8. This includes discussions on the interpolation of the signal shape and efficiency, necessary as the signal width is too narrow to utilize a full simulation. Before reaching the final remarks, chapter 9 interprets the sensitivity on the cross section as a model-dependent sensitivity on the ALP photon coupling $g_{a\gamma\gamma}$ for the case of ALPs exclusively coupling to photons. Further, this presents one potential future possibility for the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process at Belle II: the extension to partially reconstructed events, where the electron, positron, or both leptons are outside of the detector acceptance, which could increase the signal efficiency and hence strengthen the limit. After these discussions, chapter 10 concludes the thesis with the final summary.

2. Theory and Overview of Axion-Like Particles

This chapter motivates the Axion-Like Particle (ALP) and the $e^+e^- \to e^+e^-a, a \to \gamma\gamma$ search from a theoretical perspective. After a brief but more general introduction to Beyond the Standard Model (BSM) physics and two important open questions that motivate the axion and the ALP in section 2.1. Subsequently, the focus will shift towards the mathematical properties of ALPs and specifically those only coupling to gauge bosons. And explain the theoretical model used for this analysis in section 2.2.

With these discussions, section 2.3 will investigate the processes that arise from an ALP with exclusive coupling to two photons, specifically in e^+e^- collisions under the conditions at Belle II, and compare the two main diagrams. Lastly, section 2.4 will present the current status of research in ALPs with photon couplings in the MeV to GeV mass range and introduce exclusion plots as a tool to visualize the research landscape in a way that makes comparisons between experiments as well as rating the performance of a new analysis straightforward.

2.1. Motivation of Axion-Like Particles

The Standard Model (SM) of particle physics, which is the best currently known description of fundamental particles and forces, predicts most areas of particle physics with high accuracy. Yet the SM lacks descriptive power in certain areas.
One of the most well-known examples is the question of gravity and Dark Matter (DM). While gravity is a well-established fundamental force in classical physics and dominates the universe at macroscopic scales, it lacks a comprehensive quantum-field-theoretical model. Therefore, it is impossible to formulate how fundamental particles interact with gravity and how the macroscopic force is transmitted on the particle level. Closely related is the observation that different effects on astronomical scales show inconsistencies in the comparison of the visible mass in galaxies or galaxy clusters with the gravitational effects observed. The first famous observation for this is the study of rotation curves in the M31 galaxy by Vera Cooper Rubin and colleagues [9]. Later, this observation was confirmed in numerous other galaxies, and similar observations on larger scales were also made through collisions between galaxy clusters and cosmological-scale simulations. One explanation proposed for these observations is the presence of DM consisting of particles that are massive and only coupled weakly to the SM, but the final answer remains open today. These DM particles must not be electrically charged, as this would couple them to the electromagnetic force and make them observable. These requirements would be

met by neutrinos, neutral SM particles that only interact via the weak force; however, the known ones are too light to constitute the observed DM density. No more candidates for DM remain in the SM. However, theory allows expanding the SM to a more general theory by adding new particles and their interactions as long as it does not contradict current observations of the known particles. Over the past decades, numerous BSM theories have emerged, most containing a viable DM candidate that spans a vast range, from adding a single particle to duplicating the entire SM particle content or more. Yet, no hints towards a specific realization in nature, or even direct signatures from the DM particle(s) were found experimentally, despite significant efforts.

A lesser-known inconsistency in the SM can be studied with observations of neutrons: While the SM has no term that violates Charge-Parity (CP) symmetry in the strong interaction, it is allowed to add a term of the shape

$$\mathcal{L} = \bar{\theta}\frac{g_s^2}{32\pi^2}\epsilon^{\mu\nu\alpha\beta}F^a_{\mu\nu}F^a_{\alpha\beta}. \tag{2.1}$$

In this equation, $\bar{\theta}$ is the effective (free) parameter responsible for the magnitude of the strong Charge-Parity Violation (CPV), g_s is the strong force's coupling constant, and F is the gluon field strength tensor. This leads to the presence of CPV in the strong force if there is no massless quark, which there is not by current knowledge, and the parameter $\bar{\theta}$ is nonzero. The presence of CPV in Quantum Chromodynamics (QCD) would result in the neutron having a nonzero magnetic dipole moment; however, current experimental measurements constrain this moment, which translates to $\bar{\theta}$, to be very small [10]. One obvious solution is to set $\bar{\theta}$ to zero, which could have any value. As the fine-tuning of this parameter seems unnatural to many, this method is disfavoured because there is no specific reason why a parameter that could have any value is precisely zero in nature. Therefore, Peccei and Quinn provided an additional explanation by introducing a new symmetry that is broken, giving rise to the axion. This new pseudoscalar particle would naturally cancel the effects from $\bar{\theta}$ [11].
The following section provides a detailed introduction to the generalized version of the axion, the ALP, especially those which couple dominantly to photons. The remainder of the thesis will focus on searching for ALPs.

2.2. Theory of Axion-Like Particles with Photon Coupling

While theory allows couplings of ALPs to gluons and fermions as well, these models are highly constrained by measurements of flavour changing decays, which are precisely measured and mostly in agreement with the SM predictions [7, 12]. In contrast, couplings to the electroweak gauge bosons are way less constrained, as most of them need specialized search efforts and can not be reinterpreted from other measurements. To introduce ALPs with gauge couplings, this section closely follows the discussions from Dolan et al. [7].

In that case the coupling of the pseudoscalar ALP a to the SM results in

$$\mathcal{L} \supset \frac{1}{2}\partial^\mu a \partial_\mu a - \frac{1}{2}m_{\mathrm{ALP}}^2 a^2 - \frac{c_B}{4f_a}B^{\mu\nu}\tilde{B}_{\mu\nu} - \frac{c_W}{4f_a}W^{i,\mu\nu}\tilde{W}^i_{\mu\nu}, \tag{2.2}$$

with the field strengths B and W^i for the $U(1)_Y$ and $SU(2)_L$ groups respectively with the dual tensors $\tilde{B}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}B^{\alpha\beta}$ and similar for W^i. Importantly, the ALP mass m_{ALP} and the coupling to the gauge sector f_a are treated as independent parameters in this, unlike for the Peccei-Quinn axion, where there is a linear relation between those quantities.
As the Lagrangian in equation equation (2.2) contains the gauge fields B and W^i before symmetry breaking, we must evaluate it after the breaking of electroweak symmetry to extract the coupling to the physical particles $W^\pm$, Z^0, and γ we can observe. Only looking at the interactions between the ALP and the other bosons, we obtain

$$\mathcal{L} \supset -\frac{g_{a\gamma\gamma}}{4}aF_{\mu\nu}\tilde{F}^{\mu\nu} - \frac{g_{a\gamma Z}}{4}aF_{\mu\nu}\tilde{Z}^{\mu\nu} - \frac{g_{aZZ}}{4}aZ_{\mu\nu}\tilde{Z}^{\mu\nu} - \frac{g_{aWW}}{4}aW_{\mu\nu}\tilde{W}^{\mu\nu}, \tag{2.3}$$

which now contains 4 different coupling parameters g_{aXY} that are no longer independent. The relevant couplings here are those containing the photon fields, so the ALP to photon and Z^0 that translates to

$$g_{a\gamma Z} = \frac{\sin 2\theta_W}{f_a}(c_W - c_B), \tag{2.4}$$

and the coupling of one ALP to two photons

$$g_{a\gamma\gamma} = \frac{c_B \cos^2\theta_W + c_W \sin^2\theta_W}{f_a}. \tag{2.5}$$

As c_W and c_B are independent parameters, we can probe the effect of different cases of their relative magnitude. In case both are of the same size: $c_W \approx c_B$, one immediately obtains:

$$g_{a\gamma Z} \approx 0, \tag{2.6}$$

$$g_{a\gamma\gamma} \approx \frac{c_B}{f_a}, \tag{2.7}$$

which results in the coupling to two photons dominating over the ALP couplings to Zγ, ZZ, and W^+W^-. The other case, $c_W \ll c_B$, is motivated by strong constraints on c_W by measurements in flavour physics, while the constraints on c_B are less extensive. In that limit, the couplings translate to

$$g_{a\gamma\gamma} = -\frac{1}{2}\cot\theta_W g_{a\gamma Z}. \tag{2.8}$$

However, for practical uses, these cases showed no difference in the resulting physics at energy scales well below the Z^0 mass. Therefore, in the context of this thesis, it was decided to continue with the case of ALPs coupling only to photons, which will be referred to as ALP with photon coupling in the following.

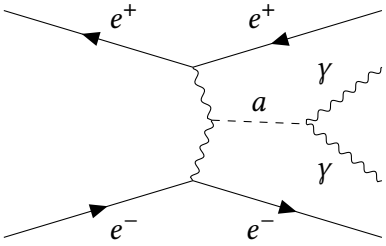
The coupling $g_{a\gamma\gamma}$ not only describes the production of the ALP, but also its decay width into two photons, and in the case of vanishing couplings to weak bosons also the lifetime τ_a, which is inversely proportional to the total decay width Γ_a. In the case of photon coupling, the width can be calculated as

$$\Gamma_a = \frac{g_{a\gamma\gamma}^2 m_{\mathrm{ALP}}^3}{64\pi}. \tag{2.9}$$

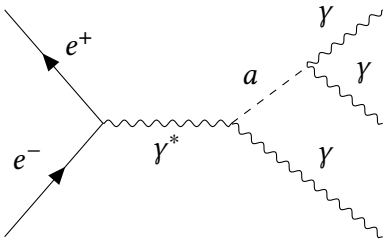
This results in a steep rise in the width and hence a decrease in lifetime for higher ALP masses, and a similar yet less steep scaling with the coupling.
All results in this work are obtained by setting a fixed $g_{a\gamma\gamma} = 10^{-3}\mathrm{GeV}^{-1}$ and calculating the width from this and the given mass.

2.3. Resulting Feynman Diagrams

As stated before, it seems a reasonable approximation to assume dominant couplings of the ALP to photons, especially as the collision energy of Belle II is approximately 10.6 GeV, which is far below the Z-mass. Furthermore, the production of the ALP should focus on the interaction of an electron and positron, as those are the particles colliding at Belle II. There are also possibilities to search for ALPs in events with B-mesons, but they are not part of this work and usually require additional couplings to W-bosons or gluons. This production and subsequent decay give rise to two leading-order diagrams: ALP-strahlung and photon fusion.
In the case of ALP-strahlung, the resulting signature is $\mathrm{e}^+\,\mathrm{e}^- \to \gamma\, a,\ a \to \gamma\gamma$. As shown in the Feynman diagram in figure 2.1 (a), the electron-positron pair produces a virtual photon in an s-channel process, which then radiates an ALP, that then can decay promptly or displaced to a pair of photons. As final state signature, this leaves three photons: two from the ALP decay and one from the ALP production vertex. This process is well studied, as an early Belle II search for ALPs decaying into diphoton resonances studied it on a limited $0.445\,\mathrm{fb}^{-1}$ dataset [13]. Additionally, Belle II is working on an update of this analysis using the current dataset [14].
The second diagram is photon-fusion, shown in figure 2.1 (b), whose signature $\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma$ this work studies. Here, the electrons do not annihilate into a single photon, but rather scatter via a photon exchange between them. This type of exchange is also referred to as t-channel. To produce the ALP, the exchanged photon between the electron and the positron can again radiate an ALP, which then also decays into two photons. In the final state, the observable particles are the photon pair from the ALP decay as well as the scattered electron-positron pair.



(a) Photon-fusion: $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$



(b) ALP-strahlung: $e^+ e^- \to \gamma\, a,\ a \to \gamma\gamma$

Figure 2.1.: The Feynman diagrams for photon-fusion (Left) and ALP-strahlung (Right). In the case of ALP-strahlung, the electron-positron pair merges into a virtual photon, which then radiates the ALP. The ALP subsequently decays into two photons, leaving a final state with three detectable photons.
For photon-fusion, the electron pair scatters via the exchange of photons, but without annihilating. The exchanged photons can again radiate an ALP that also decays into two photons. This leaves a final state with the photon pair from the ALP decay and the scattered electron-positron pair.

As described in [7], the main differences in comparison between the two diagrams are:

1. The cross section is larger for photon-fusion than for ALP-strahlung by approximately a factor of 50 for low masses and 2 for high masses, as shown in figure 2.2.

2. While the ALP-strahlung features photons that are mainly in the acceptance of the Belle II detector, the photon-fusion favours the leptons scattering at low angle and retaining most of their energy. This significantly limits the sensitivity for the fully-reconstructed case due to the geometric constraints of the Belle II track acceptance.

3. While the radiated ALP and the third photon are rather back to back in the case of ALP-strahlung, giving the ALP considerable amounts of kinetic energy, the fusion process favours ALP production at lower momenta, which results in fewer effects by elongated ALP lifetimes and resulting displacements. This also results in significantly higher opening angles between the photons from the ALP decay, making it easier to resolve the photons in the fusion case.

While cross section and ALP kinematics favour the fusion process, it was excluded from previous experimental searches due to its low signal efficiency. This could be mitigated by also studying semi-reconstructed final states where one or both leptons are outside of the detector acceptance and the signature consists of two photons from the ALP decay with one or no charged particle and missing energy from the undetected electron, positron, or both. Yet, this work will only focus on the fully reconstructed case, as the others are experimentally far more challenging and are beyond the scope of this thesis. Newer theory results [8] motivate this further, as they hint that the backgrounds for the fully resolved case might be lower than previously expected [7], making an analysis feasible.

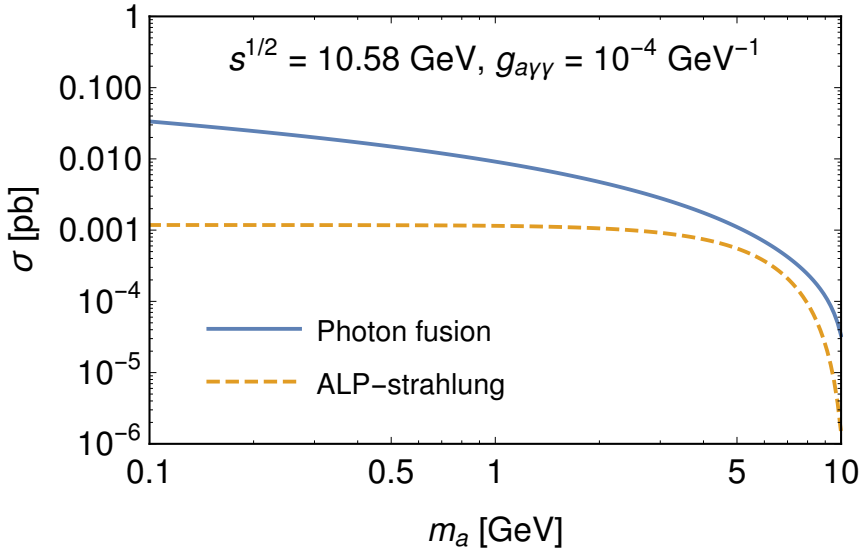


Figure 2.2.: The cross section of the photon fusion $e^+e^- \to e^+e^-a, a \to \gamma\gamma$ and ALP-strahlung $e^+e^- \to \gamma a, a \to \gamma\gamma$ process under Belle II conditions over the ALP-mass. This Plot is taken from [7].

During work on this topic, an additional contributing process to the photon fusion's final state $e^+e^-\gamma\gamma$ was found: The diagram of ALP-strahlung $e^+ e^- \to \gamma\ a,\ a \to \gamma\gamma$ but with an additional internal $\gamma \to e^+e^-$ conversion. This leads to the same final state but with entirely different kinematics. The Feynman diagram for this can be seen in figure 2.3, clearly showing more similarity with the ALP-strahlung in figure 2.1(b) than fusion in figure 2.1(a). This process was initially overlooked, as it was assumed to be suppressed with a cross section approximately four orders of magnitude below that of photon fusion. However, during later stages of the analysis, the effect of this additional signature was observed. This diagram shows considerable effects, as the expected signal efficiencies for the t-channel process are suppressed, which partially cancels the cross section suppression of the conversion leading. Therfore, between 15 to 20 % of the reconstructed events show the s-channel-like kinematics. These kinematic properties differ significantly from the t-channel kinematics: the electrons are almost parallel and back-to-back with the two photons, which also feature a low opening angle due to the high boost of the ALP. This results in an almost two-jet-like structure and very high efficiency for those events. As signal efficiency was expected to be the central limit for this search, the analysis strategy discussed in chapter 4 aims to study both diagrams at once.

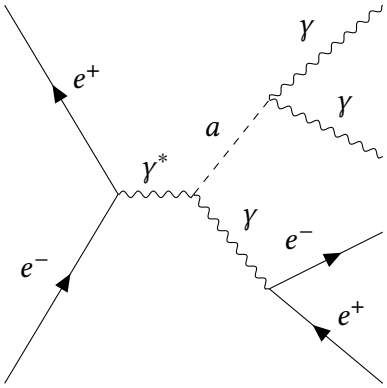


Figure 2.3.: The additional Feynman diagram that contributes to the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ signature together with figure 2.1(a).
This diagram can be tough of as ALP-Strahlung ($e^+ e^- \to \gamma a,\ a \to \gamma\gamma$, Figure 2.1(b)) with an internal $\gamma \to e^+e^-$ conversion.

2.4. Experimental Status

By now, no experimental search has seen hints of ALPs, which allows excluding the presence of signal in certain regions using statistical methods. This is usually visualized via exclusion plots, allowing a straightforward comparison of different results as explained in the next paragraphs.

As explained above, the simplest ALP models have two independent parameters: The mass of the ALP m_{ALP} and the coupling strength to the photon $g_{a\gamma\gamma}$, assuming the coupling to W- and Z-bosons is suppressed and can be neglected. This makes it possible to visualize the previous results with so-called exclusion plots, which represent previous search efforts and the parameter regions where the analyses could exclude the presence of a signal at a given confidence level, usually 95 %. These plots typically combine various search strategies, such as reinterpretations of other measurements, cosmological observations, and direct searches at dedicated experiments or multi-purpose colliders, resulting in sensitivity in different areas and exclusion curves of different shapes.

For this analysis, the relevant ALP-mass range spans from approximately $100\,\mathrm{MeV}/c^2$ to $9.5\,\mathrm{GeV}/c^2$, limited by constraints of the Belle II experiment. Reaching lower masses makes the resolution of the photons difficult, as it contains a high amount of photons with energies below 100 MeV, a region where the amount of photons radiated from the accelerator's beams is very high. On the other end, the reach at high masses is limited by the accelerator's center of mass energy, which is usually tuned at 10.58 GeV. It can not be reached fully, as the final state leptons must have momentum left to hit the calorimeter system for best accuracy.

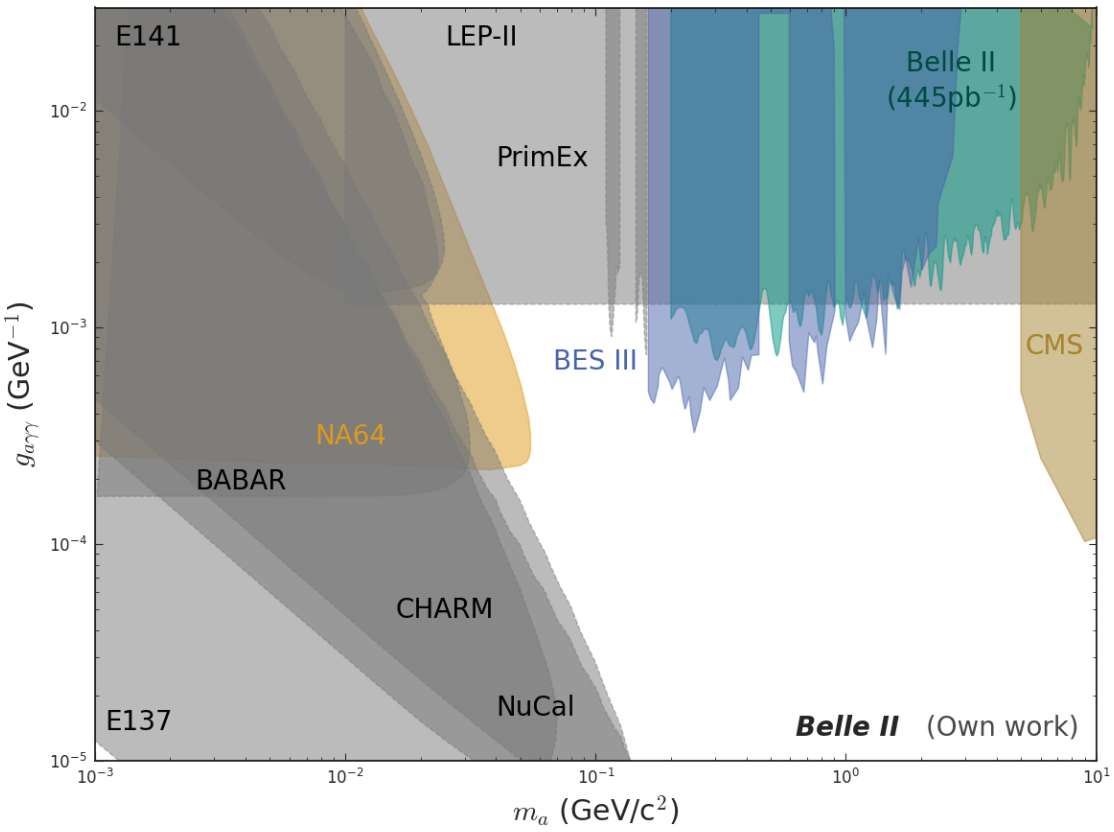


Figure 2.4.: Previous experimental limits on the ALP-photon coupling constant $g_{a\gamma\gamma}$ for ALPs with Photon Coupling over the ALP mass m_{ALP}.

This energy range is exciting as it sits in a gap between dedicated experiments at beam dumps, excluding lower masses, and the high-energy frontier experiments at the Large Hadron Collider (LHC), excluding the masses above $\approx 10\,\mathrm{GeV}/c^2$. In this region, the main results are from the early Belle II analysis of ALP-strahlung $e^+ e^- \to \gamma\, a,\ a \to \gamma\gamma$ [13] and $J/\psi \to \gamma a,\ a \to \gamma\gamma$ decays studied at the Beijing Spectrometer III (BES III) experiment [15]. Figure 2.4 displays the current status of experimental limits on ALPs with photon couplings, with the searches by Belle II and BES III highlighted, as those are the main competitors to this analysis.

3. The Belle II Experiment

This section provides an introduction to the concept of B-Factories, which are a special type of accelerator-based particle detector used to study rare processes with high detail. Beyond that, it will focus on the only currently operational B-Factory worldwide, the Belle II detector located in Tsukuba, Japan, that is used for this work. The main subsystems of the machine are introduced and explained in section 3.1 as far as needed for understanding the analysis chapter. Section 3.2 will also introduce the collaborative Belle II Analysis Software Framework (basf2), which streamlines many steps of an analysis workflow and ensures consistency and reliability of the results. Additionally, **??** introduces the centrally available datasets for standard model background Monte Carlo and beam background overlays taken from data, which this analysis uses.

3.1. The Belle II Experiment

Belle II is the successor of the Belle experiment that was in operation from 1999 to 2010 at the asymmetric KEKB electron positron collider in Tsukuba, Japan. The primary motivation for building these experiments is the measurement of properties of so-called B-mesons, especially time-dependent Charge-Parity (CP)-violation. These effects are exceptionally rare; therefore, it is mandatory to collect high amounts of up to hundreds of millions of B-meson pairs during the operation. To achieve this in a reasonable amount of time, the SuperKEKB accelerator is designed to reach previously unachieved instantaneous luminosities of up to $6 \cdot 10^{35}/(\mathrm{cm}^2\,\mathrm{s})$. The current highest recorded luminosity at Belle II is at $5.1 \cdot 10^{34}/(\mathrm{cm}^2\,\mathrm{s})$, which also resembles the world record for the highest ever achieved luminosity by any collider. With these luminosities, Belle II has collected a physics dataset with a time-integrated luminosity of $569\,\mathrm{fb}^{-1}$ since the start of data taking on 25.03.2019[1] as shown in figure 3.1.
Besides the great collision rate, SuperKEKB features asymmetric beam energies of 7 GeV electrons and 4 GeV positrons. This is typically tuned to be precisely at the mass of the $\Upsilon(4S)$ meson in the center-of-mass system. The precise energy results in resonant production of Υ mesons with a cross section of (1.110 ± 0.008) nb that is needed to produce them in sufficient rates. This specific excited state of the Υ meson is chosen as more than 96 % of them decay promptly in a pair of either charged B^+B^-- or neutral $B^0\bar{B}^0$- mesons in a coherent state. This allows the use of a technique called tagging to enhance precision in the analysis of rare decays. In the following, terms like B-mesons or $B\bar{B}$ always refer to both of those types unless stated otherwise. The asymmetry of the energies is used to boost the Υ and, consequently, the $B\bar{B}$ system, allowing for easier separation from events without the production of those mesons. Additionally, it elongates their lifetime via the

[1] Accessed 08.09.2025.

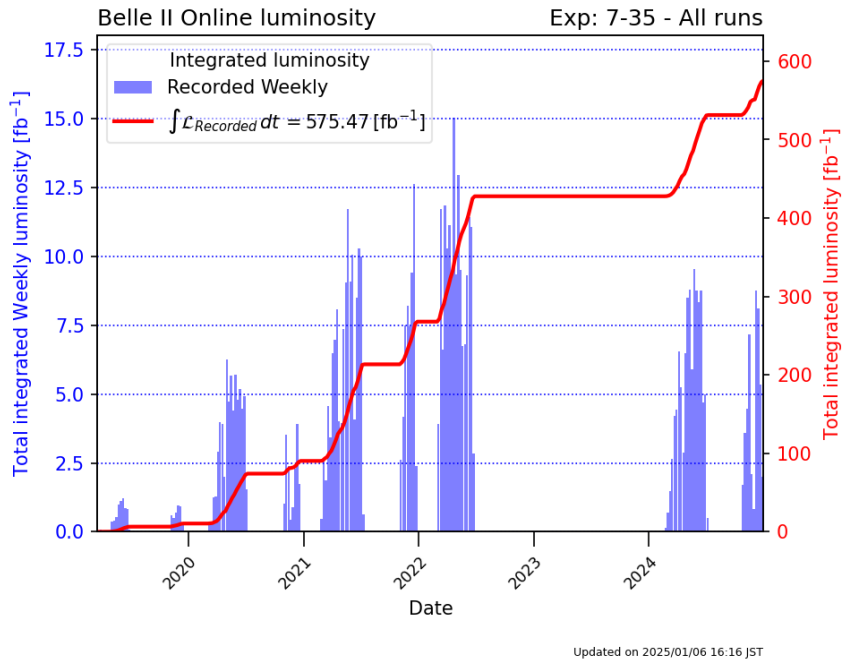


Figure 3.1.: Weekly integrated luminosity recorded by Belle II as well as the total integrated luminosity over time for the experiments up to 35 (Dec. 2024). Taken from [16].

Lorentz boost, thereby increasing the sensitivity for time-dependent analyses further. This method makes Belle II ideal for studies of B-mesons with far greater precision than possible at the previous generation of B-Factories with Belle and BaBar.

However, the SuperKEKB Belle II setup also provides a high-rate and -precision instrument for measurements outside B-physics. Especially, as the collisions without B-production exceed the B-events by orders of magnitude in cross section and with that also in rate. For example Bhabha scattering $e^+e^- \to e^+e^-(\gamma)$ has a cross section of (300 ± 3) nb[2] under Belle II conditions, which is about 270 times higher than those of Υ production. With the given data acquisition system, these rates exceed what the experiment can store and process, which is why dedicated fast electronics systems, so-called triggers, are used to filter out events that are not of interest in real-time. During the Belle to Belle II upgrades, the physics scope was extended to include interesting physics in non-B events with changes to the trigger systems that allowed recording these events with higher efficiency. These changes included adding dedicated triggers for Beyond the Standard Model (BSM) signatures like low multiplicity events, significantly extending the reach in the analysis of continuum events. This makes analyses like this one possible, and the high rate and good resolution at Belle II allow probing low couplings.

[2]Given the preselections used for the Bhabha background generator, most notably a final state lepton angle between 10 and 170° in the center-of-mass system. Without the selection, the cross section diverges [17].

3.1.1. Subdetectors and Systems

Similar to most collider-based experiments, Belle II features a cylindrical setup of tracking detectors, calorimeters, and muon systems in a strong magnetic field to measure momentum and energy of charged and neutral particles. As the focus at Belle II is highest resolution instead of energy, the setup has a few key differences compared to, for example, Compact Muon Solenoid (CMS), which will become clear when going through the different subsystems in detail, following the path that particles take when transversing from the collision point outwards through the three different tracking detectors, the Particle Identification (PID) systems, to the calorimeter and eventually the K_L- and muon system. A schematic image of the detector and its subsystems is shown in figure 3.2. For consistent reference, Belle II uses a unified coordinate system of the z-axis through the center of the detector in the direction of the high-energy e^- beam, y upwards perpendicular, and x horizontally towards the center of the accelerator. In addition to these coordinate systems, the asymmetric energy of the SuperKEKB-collider allows a clear definition of a forward and backward direction in the detector. Here, forward refers to the direction in which the high-energy electron beam points, while backward refers to the positron direction. These two terms can be used to describe the hemispheres of the detector, but more commonly are used for the endcaps, the flat base sections of the cylindrical calorimeter. The connecting lateral area of the cylinder is usually called the barrel region. These three regions of the detector will be mentioned multiple times throughout the analysis, as they bring distinct experimental features and challenges. As it is more convenient to work in spherical or cylindrical coordinates, for example, when talking about momentum directions or cluster positions, both of these coordinate systems exist as well. The azimuthal angle θ here is defined such that $\theta = 0$ means forward and $\theta = \pi$ backward.

3.1.1.1. Vertex Detectors

The Vertex Detector (VXD) is the innermost part of the tracking system, consisting of two layers of silicon pixel detectors, the Pixel Detector (PXD), at radii of 14 mm and 22 mm, and four layers of silicon strip detectors, the Silicon Vertex Detector (SVD). This system is needed, as the radiation levels close to the interaction point are high and the spatial precision must reach the µm level to distinguish particles from the interaction from those of secondary decays or beam-induced backgrounds. Additionally, the spatial precision enhances the reach for lifetime analyses and enables observations such as time-dependent CP-violation.

This approach of a pixel-to-strip tracking system is fairly common in modern detectors, aiming to compromise between the increased spatial accuracy of pixels and keeping the number of readout channels low. To further increase the performance without increasing the number of detector modules, the forward modules of SVD layers 4 to 6 are slanted, as depicted in figure 3.3, giving the VXD an acceptance between 17° and 150° without increasing the material budget significantly.

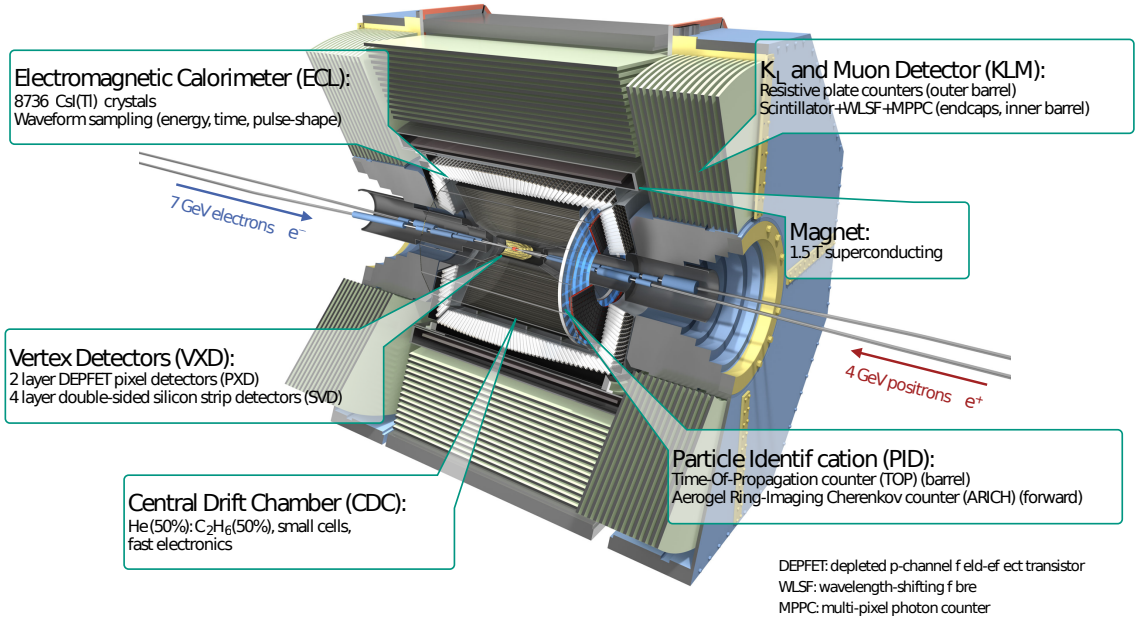


Figure 3.2.: Schematic image of the Belle II detector, its subdetectors, and the used detection techniques.
Illustration adapted from [14]. Initial Rendering from [18]

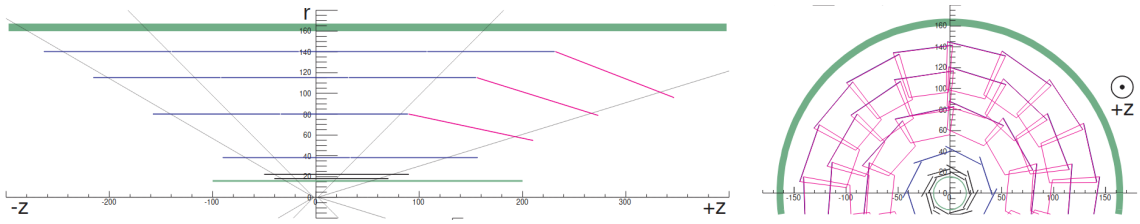
It is important to note here that during the early operation of Belle II, the PXD was not installed fully, and a new and complete pixel detector was installed only during the first long shutdown. In the early version of the PXD, only two of the 12 planned sensor ladders were mounted in the second layer, while layer 1 was already fully completed. Figure 3.4 displays a transversal view of the partial PXD setup. The conditions used for this analysis are in the initial layout with the missing modules, but the effect on the analysis result is negligible.

3.1.1.2. Central Drift Chamber

To achieve high momentum resolution, it is most beneficial to increase the radius of the tracking systems, as the relative momentum resolution $\frac{\sigma(p_T)}{p_T}$ scales for many hits following the Gluckstern formula:[21]

$$\frac{\sigma(p_T)}{p_T} \propto \frac{\sigma_x p_T}{BL^2\sqrt{N}}. \tag{3.1}$$

Where σ_x is the spatial resolution, p_T the transverse momentum component, B the magnetic flux density, N the number of measurements, and L the length of the track. However, realizing this by extending the VXD brings multiple issues. Firstly, the material budget would increase even with the thin modules, which again degrades the resolution. Secondly, the number of readout channels would be extremely high, putting higher requirements on the trigger and data acquisition system. To balance the number of channels, material budget, and resolution, the technique used for the main tracking system Central Drift



(a) Longitudinal (b) Transversal

Figure 3.3.: These figures show the longitudinal (left) and transversal (right) cut through the Belle II silicon tracking system. Black lines indicate the two layers of the pixel detector, while the blue lines are the silicon strip detector layers, with the forward-facing slanted modules highlighted in pink. This figure is taken from [19].

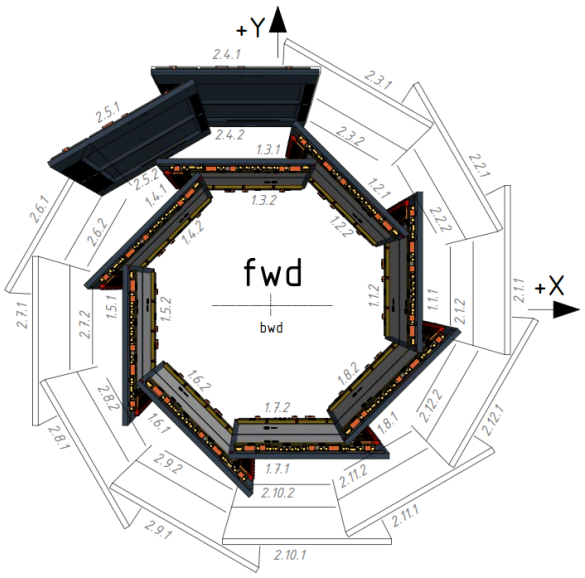


Figure 3.4.: Animation of the partial second layer of the pixel detector at the early stage of Belle II, with the missing modules marked as outlines.
The conditions used in this analysis have this partial pixel detector setup; however, a noticeable effect is not expected, as most of the tracking performance comes from the strip detectors and the drift chamber. This figure is adapted from [20].

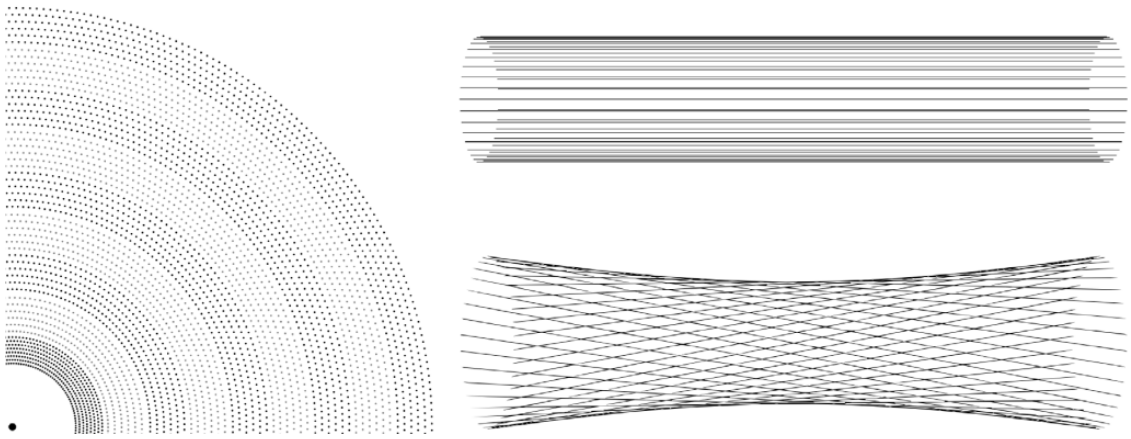


Figure 3.5.: This illustration shows the wire setup of the central drift chamber, Belle II's main tracking detector. The left picture shows a transverse view of one quadrant of the drift chamber; the right shows the different layer types: axial layers with parallel wires (top) and stereo layers with skewed wires (bottom). The skewness of the wires, exaggerated in the picture, allows for the measurement of longitudinal information. This figure is taken from [22]

Chamber (CDC) is the multi-wire proportional chamber. In general, this setup consists of many wires in a gas-filled cylindrical volume with a high voltage between them. When charged particles pass through the gas, they ionize portions of it, and the high voltage separates the charges created before they can recombine. Analog readout of the charge over time enables the resolution to be increased by measuring the relative time it takes for the charges to drift towards the hit wire, thereby constraining the path with greater resolution than the wire distance in the radial plane. Furthermore, the CDC also has longitudinal resolution that would not be possible in a design with purely parallel wires. This is achieved by using alternating layers of parallel "axial" wires and wires twisted to a hyperboloid shape in "stereo" layers, as shown in figure 3.5.
These measures combine to allow for precise tracking in the θ-range between 17° and 150° at a radius of up to 1.082 m while keeping the material budget as well as the number of readout channels low.

3.1.1.3. Particle Identification Systems

Although the tracking system can identify charged particles with the energy loss along their tracks, the forward and barrel regions contain additional PID subdetectors. Due to the boost, most particles from B-events are expected to either hit the forward or barrel region; therefore, no such detector is installed in the backward endcap, giving room for instrumentation of the CDC.
While the barrels Time of Propagation Counter (TOP) and the forward Aerogel Ring-Imaging Cherenkov detector (ARICH) both use the Cherenkov effect to measure the velocity of particles via the Cherenkov opening angle, they use different techniques to achieve this. ARICH uses two 20 mm thick layers of aerogel to create Cherenkov photons.

These photons will form cones of distinct opening angles, which are then read out using an array of square-shaped avalanche photodiodes at a fixed distance. In the readout, the circles are directly visible and allow for measurement of the angle via the radius. TOP, on the other hand, has tighter spatial constraints. Therefore, it uses thin quartz bars that contain the light, similar to an optical fiber, via reflection. Instrumenting one end of these bars with photodiodes enables the readout of the time-dependent rate of Cherenkov photons, providing the opening angle when compared to the timing information from other subdetectors.
This setup offers high accuracy PID to discriminate, for example, charged pions from kaons that are hard to distinguish experimentally otherwise. However, it is essential to note that neutral particles can neither be identified via track energy loss nor signals in TOP and ARICH as they don't produce the Cherenkov effect or tracks. Identification of those relies on the calorimeter and muon system.

3.1.1.4. Electromagnetic Calorimeter

All previous detectors focused primarily on charged particles; however, many events also contain neutral particles, especially long-lived kaons and photons from, e.g., π_0 decays. These show no signature in the tracking or PID systems; therefore, dedicated detectors are needed to measure their energy.
For this task Belle II has a Electromagnetic Calorimeter (ECL) consisting of 8736 thalium-doped cesium iodite CsI(Tl) crystals of approximately $6 \times 6\,\mathrm{cm}^2$ inner area and 30 cm length. The crystal shapes follow the asymmetry of the collision as shown in figure 3.6.
The first one is the ECL that aims at measuring the energy and position of, especially, photons and electrons via stopping them. This process generally works by forcing them to undergo cascades of Bremsstrahlung $e^{\pm} \longrightarrow e^{\pm}\gamma$ and pair conversion $\gamma \longrightarrow e^{+}e^{-}$ in high-density material, together with scintillation, until the photons are low-energetic enough to be counted with photomultipliers or photodiodes. High-energy particles, such as hadrons, photons, or electrons, typically create showers that deposit energy not only in the hit crystal but also in surrounding ones, which is referred to as a cluster. As the hadrons can also undergo nuclear hadronic interactions, the detailed shapes of the clusters differ from those expected for purely electromagnetic interactions by electrons or photons. This allows for separation between hadronic clusters and electromagnetic ones.
Belle II uses thalium-doped cesium iodite CsI(Tl) crystals to achieve a percent-level energy resolution at the cost of timing resolution with acceptable granularity. The whole calorimeter assymbly consists of 8736 crystals with approximately $6 \times 6\,\mathrm{cm}^2$ inner area and 30 cm length. Due to the asymmetric accelerator and detector design, the crystals must also be asymmetrical; in total, the calorimeter uses 98 distinct shapes to nearly fully cover the θ region between 12.4° and 155.1°. This acceptance is slightly larger than that of the tracking system, allowing showers from particles near the edges of the CDC to fully build inside the ECL acceptance without spreading into uninstrumented areas. This would harm energy resolution in the very forward and backward regions. During the upgrade from Belle to Belle II, the crystals were kept while exchanging the associated electronic components[19].

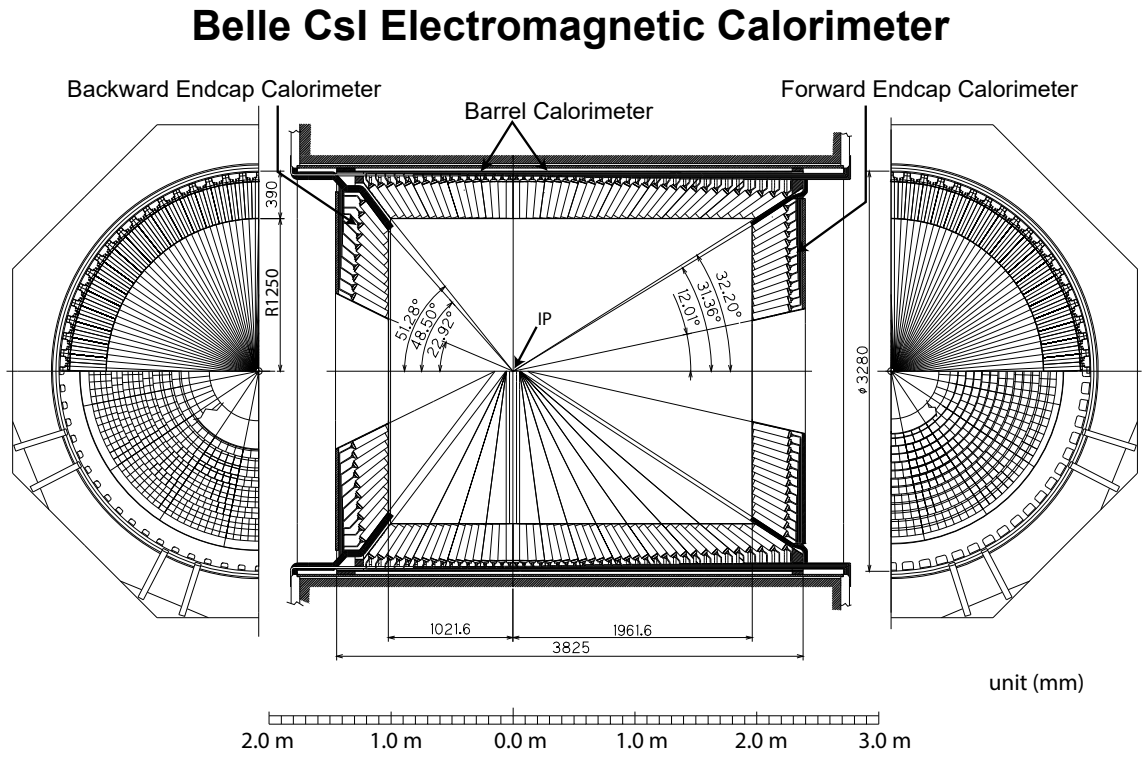


Figure 3.6.: This drawing shows the longitudinal cut through the calorimeter in the center. Additionally, a flipped view of the ϕ axis is shown in the barrel region (top left and top right), as well as the endcap crystals in ϕ (bottom left and right). One can clearly see the asymmetric layout of the detector, following the asymmetric collision energies. This figure is taken from [23].

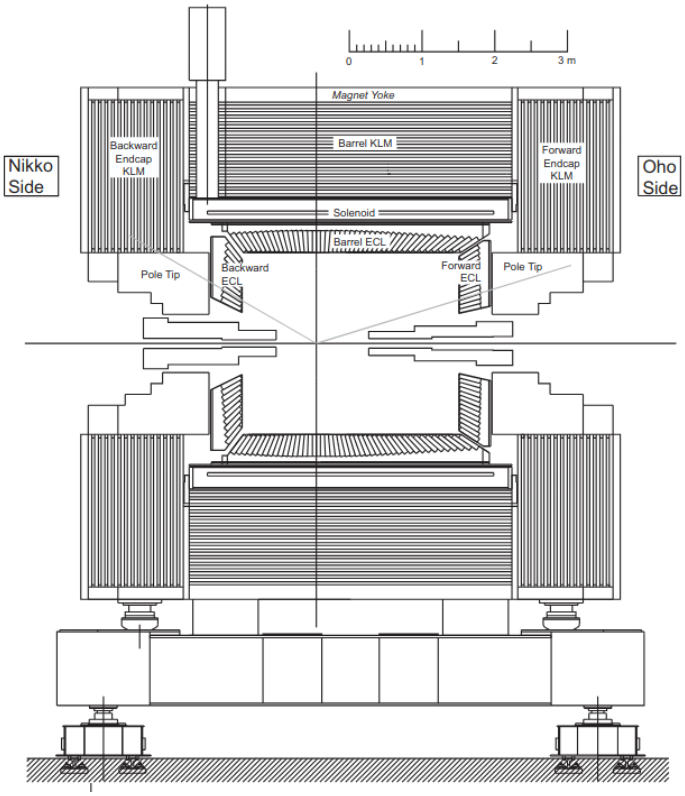


Figure 3.7.: This drawing shows a longitudinal cut through the Belle II detector. It displays the asymmetry of the electromagnetic calorimeter, and additionally, the K_L and Muon systems layers surrounding the solenoid magnet. While not directly displaying the Drift chamber, it shows the focusing magnets and shield reaching into the central white space where the tracking systems are located. These structures mainly limit the angular reach of the calorimeter and drift chamber. This figure is taken from [19]

3.1.1.5. K_L- and Muon-System

To measure the energy of unstopped particles, the K_L and Muon Detector (KLM) uses the typical sandwich structure of iron absorbers that also act as flux return of the magnet with active detector material. The Belle II KLM uses a combination of Resistive Plate Chambers (RPCs) and scintillators as a detector. This setup provides high stopping power, but the sandwich structure degrades energy resolution compared to homogeneous detectors, which are not feasible for this purpose due to price and size.
This setup provides full coverage between 20 and 155° in θ with 15 (14) detector layers in the barrel (endcaps). As figure 3.7 shows, the endcaps extend to the outer radius of the barrel region, ensuring coverage in these areas.

3.1.2. Readout and Triggers

After understanding the general setup, the working principle, and the purpose of the different subdetectors, it is left to record any useful data from the experiment. As the general number of readout channels is exceptionally high, with the PXD alone making up 8.8 million channels, and only a small fraction of events being interesting, this must be done in a well-engineered way, so as not to exceed the capacity of the data acquisition system. The crucial point here is balancing the amount of interesting physics lost with the amount of unwanted data and noise recorded. Belle II uses a two-stage system to interpret the collision data in real time and filter based on that interpretation.

The first stage is the Level One (L1) trigger, which combines the information on subdetector level to form a decision whether to keep or discard that event using Field Programmable Gate Arrays (FPGAs). However, to keep the complexity manageable, the decisions use reduced granularity and only certain subdetectors and regions. To reduce the number of readout channels, while keeping enough information for a proper decision, the L1 only uses data from the CDC, ECL, and KLM subdetectors. The main goal of the L1 trigger is the reduction of events without a collision in them and events of not interesting processes, like $e^+e^- \to \gamma\gamma$ or $e^+e^- \to e^+e^-$. This is done by combining multiple different conditions, so-called trigger lines. These can feature simple logic-based decisions like: at least one cluster of an energy above 1 GeV in the barrel region of the ECL, but they can also feature more advanced combinations. For example, Belle II also features track trigger logics that involve real-time use of a neural network to reconstruct the 3D track from only 2D inputs. Most lines also contain vetoing conditions that avoid the recording of the event, e.g., if the accelerator is currently injecting new bunches.
The individual lines each form a prediction on whether or not this line would keep or discard the event. For the high interest lines like detection of $\Upsilon(4s)$ events a keep decision always leads to keeping the event. For other lines that might trigger more often or are of lesser interest, like flagged Bhabha events, the logic contains prescales of e.g. 100 for the Bhabha trigger. This prescale means, that only 1 % of the Bhabha events are kept, while the others are discarded.

After the L1 hardware trigger decides to keep an event, it is passed to the High Level Trigger (HLT), a software trigger, which continues the decision with higher resolution. Here, dedicated Central Processing Unit (CPU) servers perform the complete reconstruction of the events in real-time, as opposed to the L1 trigger, which only runs low-granularity subdetector-level calculations on FPGAs. This also includes full-resolution clustering, as well as track finding and fitting. However, the stringent time constraints for this online reconstruction bring two main differences compared to offline: Firstly, the PXD's 8 million readout channels exceed the system's capabilities, as their datarate is one order of magnitude higher than that of all other subdetectors combined. Therefore, the PXD is excluded in the HLT and only enters after the decision. Secondly, the calibration is not performed in real-time, so the calibration settings used in online reconstruction differ from those provided for offline analyses.

This setup for the HLT serves four main purposes:

1. Physics triggers that rely on higher granularity and hence better resolution of the reconstruction,
2. Live monitoring of the data quality,
3. Providing inputs for calibrations,
4. Building regions of interest in the PXD based on extrapolating the tracks found by SVD and CDC back into the PXD.

If the trigger decision of the HLT also decides to keep an event, it is combined with the relevant readout of the PXD and passed to the data storage systems.

This two-staged setup of triggers provides sufficient reduction in data rate to make the output manageable for modern systems, while also remaining highly efficient for physics processes, and keeping smaller samples of other processes for further studies.

3.2. The Belle II Analysis Software Framework

The Belle II collaboration maintains basf2, which tackles most tasks in the workflows, from the HLT online reconstruction or Monte Carlo (MC) event generation to high-level user analysis. This ensures the consistency and correctness of the obtained results, high performance, and convenient documentation.

The modular framework consists of code written in C++ for optimal efficiency, but it also offers convenient interfaces to Python, as this is the programming language used for most user workflows.
In an analysis like this one, the user typically builds a linear chain of modules for different tasks called a basf2 path. The individual elements of this chain can be simpler tasks like boosting the generated events to the asymmetric Belle II laboratory frame, but also contain highly complex parts like the event generation itself, simulation of material interaction, or the trigger responses. Many of these steps contain some central calibration parameters that are independent of the analysis. These parameters, e.g., the beam setup or weights for analysis tools such as neural networks or decision trees for PID, are stored in a central conditions database and automatically accessed by basf2[24].
The individual modules access a common DataStore, which allows building complex chains of modules and saving the events to a ROOT file at any point [24]. The latter can sometimes be convenient to save execution time for stages where the previous outputs do not change rapidly, but are time-consuming to calculate, e.g., between the simulation and the decay reconstruction.

A typical path for an analysis might start by input files from an external event generator or by adding the module for the internal generators, which are then postprocessed, e.g., by boosting or first selections. Then, different stages of reconstruction are performed,

after which the simulation should be almost indistinguishable from actual data. From that point, the desired decay mode can be reconstructed, and selections on particle, event, or candidate level can be applied. Candidates represent the different possibilities for reconstructing a single event, e.g., via using different combinations of final-state particles. Afterwards, numerous variables for the events or particles can be calculated and stored via ROOT in a tabular format, which can be further processed. These variables encompass a range from simple physics variables, such as the momenta or energy of final-state particles, to technical variables, including low-level information on the shapes of clusters. Still, they can also cover more complex information linking multiple subdetectors. Examples for this are the shape variables of the whole event that use all reconstructed particles to describe if the event is spherical or jet-like, or the PID variables, which partially use neural networks and decision trees trained on information from all subdetectors to separate charged particles. These variables can be accessed either directly or with the use of variable collections and standard particle lists, which provide the necessary variables for a specific task and the appropriate preselections for the particle.
To further process the output, one can use ROOT, but the basf2 framework also ships various handy packages for the analysis in Python via its dependent software stack.

In summary, basf2 offers a streamlined tool with many valuable features for user analyses, which will be used heavily in the following chapter. The unification of complex parts of the workflow, like simulation or the PID tools, enables their effortless use, while ensuring consistency and providing a common language through collaboration.

3.3. Centrally Availiable Datsets

To increase consistency and efficient use of resources, Belle II has specific datasets centrally available and distributed to the collaboration. Importantly, this includes the MC background used in this analysis as well as the overlays for beam background used in the signal MC.
Additionally, the conditions in a particle physics experiment change significantly over the multiple years of operation, which influences the data taken. While some parts may slowly degrade in performance due to factors such as radiation or high-voltage damage, others may improve their performance through hardware upgrades. All these changes can potentially influence the analysis and lead to biases towards certain conditions. Therefore, the background MC and beam background samples are available on a run-by-run basis.[3]

3.3.1. Beam Background Overlays

One important reason for changing conditions at Belle II is the operation stability and delivered luminosity of the SuperKEKB accelerator. This not only influences the rate of data taking, but also the levels of background in the detector, which overlays with

[3] In Belle II terminology, a run is an up to 8 h long window of data taking without major incidents or severe changes in configurations. Experiments consist of larger groups of subsequent runs, spanning several weeks to a few months.

the actual events. Interactions between or in the beams can lead to the production of particles, mainly photons, electrons, and positrons, that are not part of the primary collision. This phenomenon is usually called beam background in the context of e^+e^- colliders as Belle II [25]. Beam background leads to additional particles, e.g., photons, in each event, not part of the initial collision, at a rate that varies drastically across different runs. To account for the beam background in simulation, so-called beam background overlays are used, which are added to the simulated events. These can be taken from a simulation or actual data.
The use of overlays from data allows for reducing the dependency on simulation and potentially increasing the MC-data-agreement, which might be beneficial in a later stage. To record these collisionless events, they are selected using a specialized trigger line [25]. Mirroring a broad range of different conditions in the analysis, MC is beneficial in previous analyses [26]. To avoid biases associated with specific operating conditions, this thesis will adopt an approach similar to this as explained in chapter 5.

3.3.2. Background Monte Carlo

As the run-dependent production, simulation, and reconstruction of Standard Model (SM) background MC is analysis independent and computationally extensive, the Mini Data Summary Table (MDST) files for the most dominant processes are centrally produced and distributed. This production contains the processes shown in table 3.1. To equalize the vastly different cross sections for the different standard model background processes, the central MC production for them has a different number of streams. While a value of 1 means the amount of MC matches the expected amount in data, a number above 1, e.g., 4 for B-events, means there is 4-times the MC compared to data. Bhabha scattering, on the other hand, has a very high cross section; hence, the produced MC dataset is downscaled to only 10 % of the expected abundance, resulting in a number of streams of 0.1.
On these samples, the detector simulation and reconstruction are already performed, and they include the proper beam background overlays and settings for the given run. This makes their usage very convenient, as they already contain everything necessary to reconstruct decays, without any exceptionally time-consuming computations.

Table 3.1.: Overview of the centrally produced background processes used in the analysis. The number of streams represents the amount of background that is simulated compared to the actual recorded data. All histograms and event counts given will compensate for this. For a better visualization, the subdominant leptonic and B-event backgrounds will be grouped in most histograms. The dashed lines indicate the used grouping.

Name	No. of Streams	Processes
$\gamma\gamma$	2.0	$e^+e^- \to \gamma\gamma(\gamma)$
e^+e^-	0.1	$e^+e^- \to e^+e^-(\gamma)$
$e^+e^-e^+e^-$	1.0	$e^+e^- \to e^+e^-e^+e^-$
$e^+e^-\mu^+\mu^-$	1.0	$e^+e^- \to e^+e^-\mu^+\mu^-$
$llXX$	1.0	$e^+e^- \to e^+e^-X^+X^-$, $X \in \{\tau, \pi, K, p\}$; $\mu^+\mu^-X^+X^-,\ X \in \{\mu, \tau\}$; $e^+e^- \to \tau^+\tau^-\tau^+\tau^-$
$\mu^+\mu^-$	1.0	$e^+e^- \to \mu^+\mu^-(\gamma)$
$\tau^+\tau^-$	4.0	$e^+e^- \to \tau^+\tau^-(\gamma)$
$q\bar{q}$	4.0	$e^+e^- \to q\bar{q},\ q \in \{u, d, s, c\}$
hhISR	1.0	$e^+e^- \to hh\gamma$, $hh \in \{\pi^+\pi^-, \pi^+\pi^-\pi^0, K^+K^-, K_S^0K_L^0\}$
$B^0\bar{B}^0$	4.0	$e^+e^- \to \Upsilon(4s) \to B^0\bar{B}^0$
$B^+\bar{B}^-$	4.0	$e^+e^- \to \Upsilon(4s) \to B^+B^-$

4. Analysis Strategy

In this analysis, I present a strategy to search for Axion-Like Particles (ALPs) produced via photon fusion and decaying visibly into two individually resolved photons at the Belle II experiment. Hence, the presence of a signal shows as a peak in the invariant diphoton mass $M_{\gamma\gamma}$ for $e^+e^- \to e^+e^-\gamma\gamma$ events over a smooth Standard Model (SM) background.

The search considers signal hypotheses in a one-dimensional parameter space, over the possible ALP-masses m_{ALP}. Due to experimental constraints the observable masses range from $0.1\,\mathrm{GeV}/c^2$ to $9.5\,\mathrm{GeV}/c^2$. Lower values are challenging, as they involve low-energy photons, which have high backgrounds from the accelerator, while higher masses leave little energy for the leptons, making their detection hard.
The width of the ALP signature is limited by the detector resolution.[1] As the final state is fully resolved, the resolution can be increased by a kinematic fit, where the final state kinematics are constrained to the initial state momentum four-vector. The kinematic fit reduces the measured signal width to the order of roughly $10\,\mathrm{MeV}/c^2$.
While non-zero lifetimes for the ALP are allowed, this search considers only prompt decays. Investigating only prompt decays is further motivated by the results of Dolan et al. [7], which show that displacement becomes significant only for very small couplings $g_{a\gamma\gamma} < 10^{-4}\,\mathrm{GeV}^{-1}$.

The analysis is confronted with two major challenges, which lead to the exclusion of the photon fusion final states from previous ALP searches at Belle II. Firstly, the t-channel kinematics of the diagram feature low momentum transfer to the final ALP and hence, the final state leptons are mostly not scattered into the detector acceptance and escape undetected. As the analysis only studies the fully reconstructed case, the signal acceptance from the geometric constraints and the θ acceptance of the Belle II detector is limited to the 0.1 to 2 % level. Secondly, the search features background from processes with very high cross sections, most notably (double-)radiative Bhabha scattering $e^+e^- \to e^+e^-(\gamma)$. These lead to high numbers of background events. Therefore, the analysis strategy is designed to reduce these backgrounds to manageable rates while keeping the signal efficiency loss acceptable.

[1] The theoretical width is calculated by the Monte Carlo event Generator Madgraph5 during signal generation, and it is found to be orders of magnitude smaller than the experimental width resolution.

To achieve good background rejection without compromising the signal efficiency further, I will present a strategy of carefully engineered selections to target specific backgrounds, combined with a neural network-based event classifier. The classifier is responsible for most of the background reduction using a variety of different kinematic and geometric variables to separate the signal from the background.

As the low signal width makes a full simulation of masses spaced with the minimal signal width unfeasible, the analysis utilizes interpolation of the signal shape as well as signal efficiency. The interpolation allows us to describe the signal for arbitrary mass values in the allowed region using smooth functions. It thus enables the calculation of sensitivity on a fine mass grid.
The sensitivity calculation uses a fit to the SM background to extract the signal strength. Subsequently, the expected sensitivity on the $e^+e^- \to e^+e^-a, a \to \gamma\gamma$ cross section is extracted assuming observing the expected background using the asymptotic approximation.

Additionally, the expected sensitivity on the cross section is converted into a model-dependent upper limit on the ALP photon coupling $g_{a\gamma\gamma}$, assuming this coupling is the only non-negligible one. The conversion of the results into $g_{a\gamma\gamma}$ allows comparison of the expected sensitivity with the constraints from other experimental searches for ALPs with photon coupling.

While this provides an overview of the complete analysis workflow, the following chapters introduce the individual steps in greater detail.

5. Dataset

This analysis utilizes centrally produced run-dependent background samples, as introduced in section 3.3.2, together with privately generated signal. Section 5.1 explains the setup and procedure used for the signal Monte Carlo (MC) simulation and reconstruction in detail, while section 5.2 explains the background dataset.

5.1. Signal

While the background is centrally produced, reconstructed, and provided to the entire collaboration, the signal for this analysis is produced privately. The production of every run, weighted to the luminosity, would be ideal, but it is far too costly in terms of computation time, as the MC is produced for many Axion-Like Particle (ALP) mass hypotheses. As previous analyses [26] revealed effects from changing experimental conditions between runs, especially with the beam background level, this study employs a collection of six runs of signal Monte Carlo, shown in table 5.1. These runs cover most of the possible range of the beam background levels, thereby avoiding bias in the analysis. If a bias occurs is checked at multiple stages of the analysis.

5.1.1. Signal Event Generation

While the theory model for the ALP with photon coupling predicts the properties of the different final states and diagrams, given the input parameters like collision energy, ALP mass, etc., this is not directly usable by experimentalists. The mathematical description must first be expressed as a model that the event generators can understand. For the case at hand, the generator used for signal MC is MadGraph5 aMC@NLO, which is included in Belle II Analysis Software Framework (basf2) [1]. As there are previous searches for the identical ALP model with photon couplings, the model also exists and is already included in basf2. It is adapted from a search for ALPs in peripheral heavy-ion collisions at the Large Hadron Collider (LHC) [27]. The generator calculates the resulting momentum and energy distributions and generates events by sampling from those distributions.

To better match the physics processes that are expected from actual data, MadGraph offers the possibility to automatically include the effects of Initial State Radiation (ISR). In this process, the beams typically radiate soft photons, which lowers the collision energy and might shift the event slightly. However, it is essential to note that due to technical limitations of MadGraph, the calculation does not give the photons as MC-particles and only shifts the kinematics of the other particles. The missing ISR photons are expected to have little to no impact, as most of them are very low energy and collinear with the

Table 5.1.: Runs used as signal Monte Carlo in chronological order. These runs, as they feature a high integrated luminosity and cover a broad range of beam background conditions. The background from these runs is used with the given luminosities, totaling 2.159/fb. Additionally, the background of the remaining runs of experiment 26 is used for evaluation, which is 52.082/fb.

Experiment	Run	Integrated Luminosity in 1/fb
22	30	0.254
22	468	0.266
24	985	0.609
24	2176	0.303
26	898	0.376
26	1485	0.351

Table 5.2.: The number of generated events per mass for the different mass ranges. As the signal efficiency changes significantly with the ALP mass, it is necessary to add this split to ensure that the available number of events is roughly in the same order of magnitude. Otherwise, the over- or undersampling might lead to issues when proceeding with the selection.

ALP mass range in GeV	Number of generated events per mass
$[0.1, 1.0[$	$12.5 \cdot 10^6$
$[1.0, 5.0]$	$2.5 \cdot 10^6$
$]5.0, 9.5]$	$1 \cdot 10^6$

beams. Therefore, most of them would not contribute even if they were fully simulated as shown in [14] for the ALP-strahlung analysis.

The signal generation utilizes a fixed coupling of $g_{a\gamma\gamma} = 1 \cdot 10^{-3}/\text{GeV}$, with the beam energy set to the run's measured value in a symmetric collision, as basf2 will later boost the events into the proper Belle II laboratory-frame. To guarantee the stable operation of the ISR correction, the beam density functions were set to $e^{\pm}$ and the renormalization and factorization scales were fixed to the set beam energy.[1] The signal generation happens independently on a run-by-run basis for the runs seen in table 5.1 and with a mass-dependent number of events per mass seen in table 5.2. Why this split was chosen is discussed further in section 5.1.3

[1] During the early phase of this study, an issue when using the ISR-correction for the photon-fusion diagram was observed, which is presented in appendix A.1

5.1.2. Signal Event Simulation and Reconstruction

The events generated by MadGraph already contain all of the theoretical expectations on the ALP, but lack the experimental side; therefore, a chain of basf2 modules is used to simulate all experimental effects.

Firstly, the collisions at Belle II are asymmetric as discussed previously, but MadGraph uses symmetrical energies. Therefore, the events must be boosted to the right reference frame to match the expected kinematics at Belle II. Afterward, a custom module, the VBF-tagger, performs a pre-selection on the lepton angles, which is further discussed in section 5.1.3, to discard events outside the detector's acceptance as early as possible, thereby saving computational time. The selection only leaves events where both photons and leptons are inside the acceptance of the detector, and therefore detectable.

In the next step, the complete simulation of all physical processes the particles undergo in the detector is performed. This simulation includes curving the trajectories of charged particles inside the magnetic field, and a detailed simulation of material interactions using Geant4.[28] Geant4 uses MC methods to calculate statistical physical effects from particles passing through the numerous active and passive materials in the detector. These effects aim to cover all processes that real particles can undergo when passing through the detector, encompassing a wide range of phenomena such as pair conversion, Bremsstrahlung or scattering, to transition and Cherenkov radiation, as well as nuclear interactions. Most of these processes produce additional particles that are also detectable. These secondary particles are simulated using the same rules.

By now, the event only consists of the four particles that MadGraph generated, and the secondaries they produced in the previous step. But as mentioned before, in the real detector, there are numerous additional background particles arising from beam interactions. To add these, the generated events are combined with so-called beam background overlays taken from data of the runs in table 5.1 as described in section 3.3.1. In the next stage, the trigger simulation is performed. This reproduces the low-granularity subdetector-level reconstruction that the Level One (L1) trigger performs on hardware to simulate whether the event would be kept by the L1 trigger or not. However, for MC studies, also events that would be removed by the trigger in data are kept on MC level. The decisions of the global logic, and individual trigger lines are attached to the events and stored for later use.

Afterwards, the actual reconstruction of objects like tracks and clusters is performed using the elaborate reconstruction algorithms implemented in basf2. The two most important parts of the reconstruction for this analysis are the track finding and fitting, together with the Electromagnetic Calorimeter (ECL) clustering. For the tracks, the Central Drift Chamber (CDC) track finding algorithms try to search for tracks in the CDC first by finding them using two complementary approaches: a global Legendre transformation-based one, and an additional one using cellular automata to detect short and displaced tracks. The found tracks are afterwards fitted with the curve of a particle curving in a magnetic field and losing energy, to get the momentum of the track. The complete explanation of these algorithms is beyond the scope of this work, but is explained in detail by Bertacchi et al. in [22]. For the ECL clusters, the algorithm works by first identifying local maxima that measured more than 10 MeV energy, and then iteratively adding the neighboring crystals with energy deposition. Kou et al. describe the algorithm in [25]. The information on

the found track and cluster objects is again attached to the events. These will also be the general lowest level of information accessible later, as detailed hit-level information is not interesting for analysis and would lead to excessively large file sizes. Therefore, the low-level technical information is discarded when saving the files later.
To utilize the potential of the MC simulation fully, the found objects are matched to the generated primary and secondary MC particles following certain rules. This step, called MC-matching, allows checking whether a particle is the true one that was generated or is coming from any other source, e.g., beam background, later in the analysis. [29] describes the used algorithm for MC-matching at Belle II.
After these steps, the resulting objects and additional information are stored in so-called Mini Data Summary Table (MDST) files using ROOT [30]. These files are identical to what is stored for the centrally produced background and will enter the same decay reconstruction described in chapter 6.

5.1.3. Dealing with Low Acceptances

As mentioned multiple times in the previous discussions, the whole analysis of $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ is challenged by the low signal efficiency due to geometric constraints of the Belle II experiment. There are two naive approaches to mitigate this issue. First, one could simulate more events and reconstruct them; the decay reconstruction introduced in chapter 6 will remove the events that are not suitable. The second possibility is to introduce a preselection in the MC generator. However, both of these solutions have issues that make them hard to use. In the first case, the high number of events leads to very long runtimes for the simulation, which are infeasible and would extensively use the available disk space with events that will not be considered anyway. The second solution would alter the cross section calculated by MadGraph, which is needed for the final sensitivity calculation.

The proposed solution tries to combine the advantages of both approaches while mitigating most of the adverse effects. As the event generation in MadGraph is relatively fast, and the files are small, producing a large number of events is feasible here. These events are then read into the basf2 reconstruction and boosted to the Belle II lab-frame, which is also fast with less than 1 ms per event. Now the events are passed through a custom basf2 module, which reads the lepton charge and angle and decides whether electron, positron, none, or both are inside the CDC acceptance. An integer flag then encodes the result of this check for each event. Based on the desired case, a selection on the integer code can be applied, which removes all unwanted events.
Afterward, the events can be passed to the remaining detector simulation, which takes up to 500 ms per event. The explained preselection method enables the production of large numbers of events while keeping execution time within an acceptable range. Furthermore, since no selections are applied during event generation, the calculated cross section from the MC generator is accurate.

With the tagger module, the number of events to generate was chosen manually in a way that results in $O(10^4)$ events after the selections. Due to the high change of reconstruction efficiency with mass, the mass range is split into three regions as shown in table 5.2. The split ensures that all masses have the same order of magnitude in the number of reconstructed events without significantly over- or underrepresented masses.[2]

5.2. Background

As described in section 3.3.2, the analysis uses centrally produced backgrounds, so the simulation and reconstruction of the files are already done, and they can enter the decay reconstruction.
Due to technical limitations, this analysis considers only a $\int L dt \approx 54\,\text{fb}^{-1}$ subset of the Belle II dataset, which represents $\approx 10\,\%$ of the currently recorded dataset. The background dataset consists of the six runs shown in table 5.1 together with the remaining runs of experiment 26. The limited luminosity is used as the main background processes $e^+e^- \to e^+e^-(\gamma)$ and $e^+e^- \to \gamma\gamma(\gamma)$ feature high cross sections, resulting in large file sizes. Processing a higher luminosity would not be feasible on the current setup and would require a technical restructuring of the workflow.[3]
To keep the version of the basf2 up to date and to account for changes, the production of background MC is repeated every few years in MC-production campaigns. At the point that this analysis started, the latest campaign was MC15, but during work, a new one, MC16, was finished. However, the new MC campaign brought significant changes to the file structure, necessitating a restructuring of the workflow. Therefore, the analysis is still using MC15 and is not updated to MC16.

Another group of background components is peaking backgrounds, which are standard model processes that mimic the signal signature. This is the case for processes that produce a neutral hadron $h^0 \in \{\pi_0, \eta, \eta'\}$ via photon fusion in $e^+e^- \to e^+e^-h^0$, that subsequently decays into two photons. These events appear almost identical to the signal of an ALP with the same mass as those hadrons. For a better understanding of the peaking backgrounds, most analyses also simulate larger samples of these processes. However, this approach will not be used here, as the processes are only dominant in the mass region in the close vicinity of the mass peak m_{h^0}, but show little effect outside. This is justified as these regions, shown in table 5.3, will be excluded in the scan, and therefore the analysis will not be sensitive to them.

[2] Madgraph has a hard limit on the number of generated events per execution. Appendix A.1 presents the way used to mitigate this.

[3] The main limit here is the local storage available at GridKA, as the $\approx 54.2\,\text{fb}^{-1}$ dataset already is around 25 TB input file size, and the number of individual files for processes that show practically no contribution to the background, e.g. B-events is large. Both could be removed by either developing a skim for this analysis, which takes time to develop and process, or by moving to other computing resources, e.g., the Belle II grid at the cost of an increased turnaround time.

Table 5.3.: Regions excluded from the signal scan over the invariant di-photon-mass $m_{\gamma\gamma}$ due to irreducible standard model background from $e^+e^- \to e^+e^-h^0, h^0 \to \gamma\gamma$ with $h^0 \in \pi^0, \eta^0, \eta'$ and the mass of those hadrons.

Hadron h^0	Mass $m(h^0)$ in GeV/c^2	Exclusion region in GeV/c^2
π^0	0.135	[0.09, 0.16]
η^0	0.548	[0.45, 0.63]
η'	0.958	[0.91, 1.01]

6. Decay Reconstruction and Postprocessing

Up to this point, the reconstruction is unaware of the signal signature being studied. To introduce the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ signature, the decay reconstruction can read the reconstructed files, and combine the information from the Mini Data Summary Tables (MDSTs) first to map the reconstructed objects to the desired observable particles, and then build decays from that. The decay reconstruction also enables the calculation of physics variables in a tabular format, which is understandable for standard data analysis tools. These can then be used to calculate further variables or conveniently apply selections.

The reconstruction of final state particles, decays, and physics variables from the reconstructed objects in the MDST files is implemented in Belle II Analysis Software Framework (basf2). The process begins with detectable final-state particles. These must be sufficiently interacting and long-lived like electrons, muons, photons, protons, and some mesons. Other particles, such as B mesons or τ leptons, decay too quickly to leave direct, visible signatures and are reconstructed only from their decay products. To reconstruct the final-state particles, the reconstructed objects, like tracks or clusters in the MDST files, are assigned to particle lists using specific criteria. However, these lists rely only on coarse physical properties and do not guarantee that they will contain only the desired particles. For example, if one requests the list of all photons, the algorithm will return all clusters that do not have a track attached to them and use the hypothesis "photon" for all subsequent calculations on that list. To remove the misreconstructed particles, one must apply different stages of selection, which can begin as early as when building the particle lists.

As the final state particles are not necessarily produced instantly upon collision, but undergo a chain of decays beforehand, the particle lists can be combined to form a decay, which yields the list of the newly constructed mother particle. For example, the Axion-Like Particle (ALP) list is built by combining two photons from the photon list, which is filled with all photons in an event that pass the selection shown in table 6.1. The final-state photons must be within the tracker's angular acceptance to be considered for the analysis. As explained in section 3.1, the Central Drift Chamber (CDC) acceptance is narrower than the Electromagnetic Calorimeter (ECL) acceptance, which would be the detector in principle responsible for photons. However, the photons must be within the tighter track acceptance to avoid misidentifying charged particles, such as electrons, as photons. If those are outside the CDC but inside the ECL acceptance, their tracks cannot be measured due to the lack of a tracking detector in this region. Therefore, there is no reliable way to discriminate between electrons and photons in that region. Additionally, all photons must fulfill a minimal energy requirement of $E > 0.1\,\text{GeV}$. This reduces beam background, as the

beam interactions produce many low-energetic photons. However, the energy selection also limits the sensitivity for the lowest ALP masses, where low-energy photons also occur frequently in the signal. Additionally, the timing fit of the cluster must be successful, which removes incorrect clusters formed, for example, by noise and reconstruction errors. However, as there can be multiple photons in the list per event, there will eventually also be many possibilities to build an ALP from photon pairs. Those combinations are called candidates, and usually only one of them is true. To remove unphysical candidates, the ALP candidate must pass a minimal transverse momentum $p_t > 0.1\,\mathrm{GeV}/c$ and energy $E > 0.5\,\mathrm{GeV}$ selection. These are needed to remove unphysical candidates from the combination of low-energy photons, e.g., from the beam background.
Combining lists to a decay can be repeated multiple times, which is necessary, for example, in B-physics where the B meson usually undergoes many decay steps before the particles become detectable. However, since the decay process for this analysis is straightforward, the reconstruction chain is relatively short, comprising only two stages. In a first stage, the ALP is reconstructed from the two photons as mentioned, and in a subsequent step, the "decay" of the beam into an e^+e^- pair and the ALP is reconstructed.
The final-state signal e^+e^--pair is from the initial collision, hence the tracks must roughly extrapolate to the interaction point, which is expressed by the $\mathrm{d}r < 3\,\mathrm{cm}$ and $|\mathrm{d}z| < 3\,\mathrm{cm}$ requirements on the minimal radial ($\mathrm{d}r$) and axial ($\mathrm{d}z$) distance of the track to the interaction point. This reduces the number of beam background tracks and reconstruction errors, as these are typically outside the window and further away from the interaction point. Additionally, the Bhabha background contains a considerable amount of backscattered leptons, whereas the signal does not. To remove them, a charge-dependent angle selection on the lepton is applied. These remove the events where the electron was scattered into the backward direction or the positron into the forward direction. While the leptons could be described entirely by the tracking, they are required to be matched to a cluster, which increases the accuracy by the ECL measurement. This requirement helps in Particle Identification (PID) and using consistent variables throughout all signal events. At the same time, it lowers the efficiency for high ALP masses, as these feature low energetic leptons, which curl too strongly in the magnetic field and do not reach the calorimeter.
The reconstructed lepton pair is combined with the reconstructed ALP to a candidate if it passes a missing mass selection $M^2_{\mathrm{Recoil}} < 5\,(\mathrm{GeV}/c^2)^2$, which avoids partial reconstruction of more complex events. Additionally, information on the whole event, not just the reconstructed candidate, can be used for further selections, e.g., the number of found tracks passing the same distance criteria as the signal tracks. This selection again removes larger events by only allowing between two and three tracks from the origin per event. The loser selection provides flexibility for an additional track that is not reconstructed. As additional tracks can also arise from beam backgrounds or mistakes in the track finding, this should not lead to the removal of a whole event. The looser selection is especially beneficial, as previous analyses have shown that strict track requirements lead to a strong dependence of the signal efficiency on background levels, introducing a run-dependent variation in the signal efficiency [26, 14]. The loose selection tries to mitigate this bias. Additionally, a simultaneity selection requires the ECL clusters to be measured in a certain time window around the average cluster time, which should further suppress beam background effects. The beam background clusters are not necessarily from the same

Table 6.1.: List of selections applied during reconstruction. Most of them aim at reducing the number of candidates that are wrongly combined by excluding very low energetic clusters, etc., or events that are certainly not of interest, e.g., by the recoil mass. The Variables and their definitions, as well as the motivation for the selections, are explained in the text.

Variable describtion	Criterium
Photons inside tracker acceptance	inCDCacceptance == 1
Minimal photon energy	$E_\gamma > 0.1\,\text{GeV}$
Photon timing fit successfull	Cluster Timing Error $< 1 \cdot 10^6$ ns
Letpon point of closest approach radial	$dr_{e^\pm} < 1\,\text{cm}$
Letpon point of closest approach longitudinal	$\lvert dz_{e^\pm}\rvert < 3\,\text{cm}$
Lepton tracks matched to cluster	clusterTrackMatch == 1
Positron not backscattered	$\theta_{e^+} > 0.52$
Electron not backscattered	$\theta_{e^-} < 2.44$
ALP candidate minimal transverse momentum	$p_{t,ALP} > 0.1\,\text{GeV}/c$
ALP candidate minimal Energy	$E_{ALP} > 0.5\,\text{GeV}$
Event missing mass	$m^2_{Recoil} < 5(\text{GeVcc})^2$
Cluster Simultainity	< 5 ns
Number of good tracks range	$nCT = [2, 3]$

bunch crossing or interaction, hence most of the background photons show significant deviations in timing from the proper photons.
Table 6.1 shows a summary of the used selection criteria in reconstruction.

With the fully reconstructed decay, basf2 can calculate numerous variables of the reconstructed objects using the hypothesis of the given particle. For the later analysis, there is a vast range of interesting variables, starting from the physics variables like the momentum components, energies, and angles of the final and intermediate particles. More technical variables like the number of hits in the CDC or ECL subdetectors are also used, e.g., to study the beam background. For more complex universal workflows, like the PID, suppression of beam background or fake photons, the variables also include the outputs of intricate MultiVariate Analysis (MVA) algorithms like decision-trees or neural networks, that combine variables from multiple subdetectors to a single numerical output.

The decay reconstruction also includes two other crucial tasks for the analysis: a kinematic fit and a vertex fit. The vertex fit attempts to determine the common origin position of two or more charged particles by fitting their tracks to a common intersection point using the TreeFit algorithm [31]. In the signal events, the electron-positron pair originates from the interaction point and therefore should not be displaced. However, one major background source is $e^+e^- \to \gamma\gamma(\gamma)$ where one of the photons performs a pair conversion $\gamma \to e^+e^-$ to produce the final state leptons. As the pair conversion needs material interaction, the earliest possible vertex is the beam pipe, which is a few millimeters away from the

interaction point. With the vertex fit, the difference becomes visible in most events in the fitted displacement distance, and a selection on the displacement, discussed in chapter 7, is applied.
Kinematic fitting, on the other hand, is a powerful tool at lepton colliders that utilizes the well-known initial state. As the leptons are elementary particles, the collision energy and momentum are known to great precision. By conservation laws, this also constrains the momentum of fully reconstructed final states. The kinematic fit tries to adjust the measured final state kinematics to match the initial state using the OrcaKinFit algorithm [32]. This significantly improves the resolution, especially on the scan-variable $M_{\gamma\gamma}$ as shown in figure 6.1 for four exemplary signal masses.
The adjustment modifies the kinematic variables for the particles while also preserving the initial values. However, the adjustment of variables leads to ambiguity in some variables if the value before or after the fit is used. The ambiguity is resolved as variables before the fit are always denoted with a "BF". Hence, if a variable has no such annotation, it either means the variable's value after fit, or that the variable is not affected.

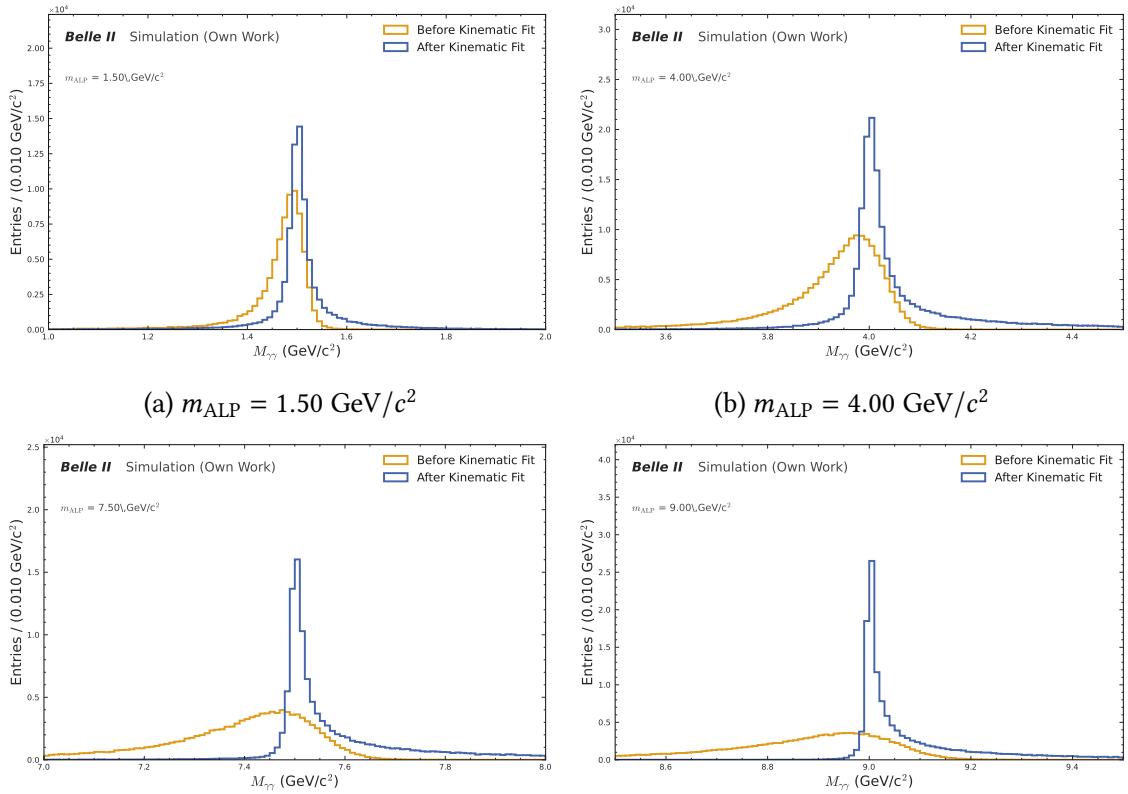
From this point, the analysis continues without the use of basf2 and instead utilizes parquet files [33], along with Pandas [34] and Polars [35]. While pandas is a well-established framework for data analysis, Polars is relatively new. It is specialized for lazy evaluation of operations on data files and can handle files that exceed the memory limitations of the system in size using a technique called streaming.
Additionally, the overhead of remaining candidates should be removed by applying a final selection that enforces one candidate per event, the best candidate selection.

The reconstruction still allows for multiple candidates per event, which must be removed for further analysis by the so-called best candidate selection. In the proper signal signature, the four reconstructed final particles should be all particles produced in the collision; hence, there should not be any additional or missing final state particles showing as missing or extra momentum or energy. One variable that expresses this is the invariant mass of the four-body system $M^{BF}_{e^+e^-\gamma\gamma}$. For proper $e^+e^- \rightarrow e^+e^-\gamma\gamma$ events, the result should be the same as the beam energy E_{Beam}. $M^{BF}_{e^+e^-\gamma\gamma} > E_{\text{Beam}}$ happens in cases when reconstructed particles are not part of the initial collision, e.g., from beam background, and $M^{BF}_{e^+e^-\gamma\gamma} < E_{\text{Beam}}$ means missing particles in the reconstruction. Therefore, this fact is used as a candidate selection criterion, and the candidate with the lowest absolute difference between beam energy and invariant mass is kept, while the others are removed. It is essential to note that the selection relies on using the mass before the kinematic fit, as the fitting process shifts the 4-body mass to match the beam energy exactly.

The signal efficiency of the reconstructed events $n_{\text{Rec.}}$ relative to the number of generated events $n_{\text{Gen.}}$

$$\epsilon_{\text{Sig.}} = \frac{n_{\text{Rec.}}}{n_{\text{Gen.}}} \tag{6.1}$$

after the selection is shown in figure 6.3 for the 6 runs of signal Monte Carlo (MC) used. The low values $< 2.5\,\%$ are dominated by the geometric acceptance of the Belle II experiment



(a) $m_{\mathrm{ALP}} = 1.50\ \mathrm{GeV}/c^2$

(b) $m_{\mathrm{ALP}} = 4.00\ \mathrm{GeV}/c^2$

(c) $m_{\mathrm{ALP}} = 7.50\ \mathrm{GeV}/c^2$

(d) $m_{\mathrm{ALP}} = 9.00\ \mathrm{GeV}/c^2$

Figure 6.1.: These plots show the impact of the kinematic fit on the reconstructed signal shape over the scan variable $m_{\gamma\gamma}$. While the effect for lower masses like $1.5\ \mathrm{GeV}/c^2$ in the top left is minimal, it becomes more visible in the intermediate mass range, e.g., $4.0\ \mathrm{GeV}/c^2$ (top right). It is evident that the fit significantly reduces the left tail and narrows the signal peak; however, it creates a right tail. For higher masses (bottom), these effects become even more pronounced. Additionally, the reduced bias on the peak position becomes directly visible.

for the final state leptons. Throughout most of the mass range, the efficiency shows a linear behaviour rising between 0.2 and 2.5 % with the ALP mass. The highest and lowest masses, however, show a steeper drop below the linear behaviour, arising primarily from the photon energy (low masses) and lepton cluster (high masses) requirements. Additionally, the efficiencies do not feature a strong run dependence, which is beneficial for the later analysis.

Figure 6.2 shows the background spectrum resulting in the scan-variable $M_{\gamma\gamma}$ at this analysis stage. The figure shows that background from Bhabha-scattering $e^+e^- \rightarrow e^+e^-(\gamma)$ is the most dominant process with noticeable contributions from diphoton processes $e^+e^- \rightarrow \gamma\gamma(\gamma)$ and other leptonic events.

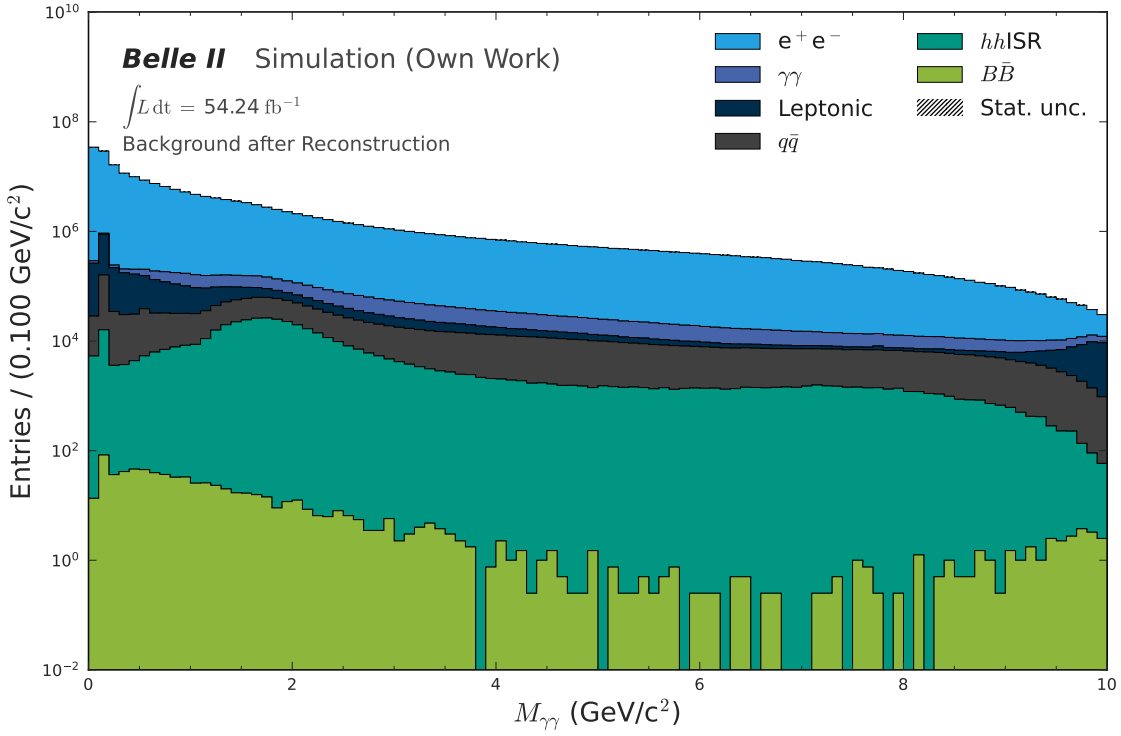


Figure 6.2.: The background distribution in the scan variable $M_{\gamma\gamma}$ for all simulated background processes using the full available luminosity of $54.24\,\mathrm{fb}^{-1}$. To better visualize the individual components, the y-axis is logarithmically scaled. This shows that the dominant backgrounds are arising from Bhabha scattering and diphoton processes, while hadronic and heavy lepton events are subdominant.

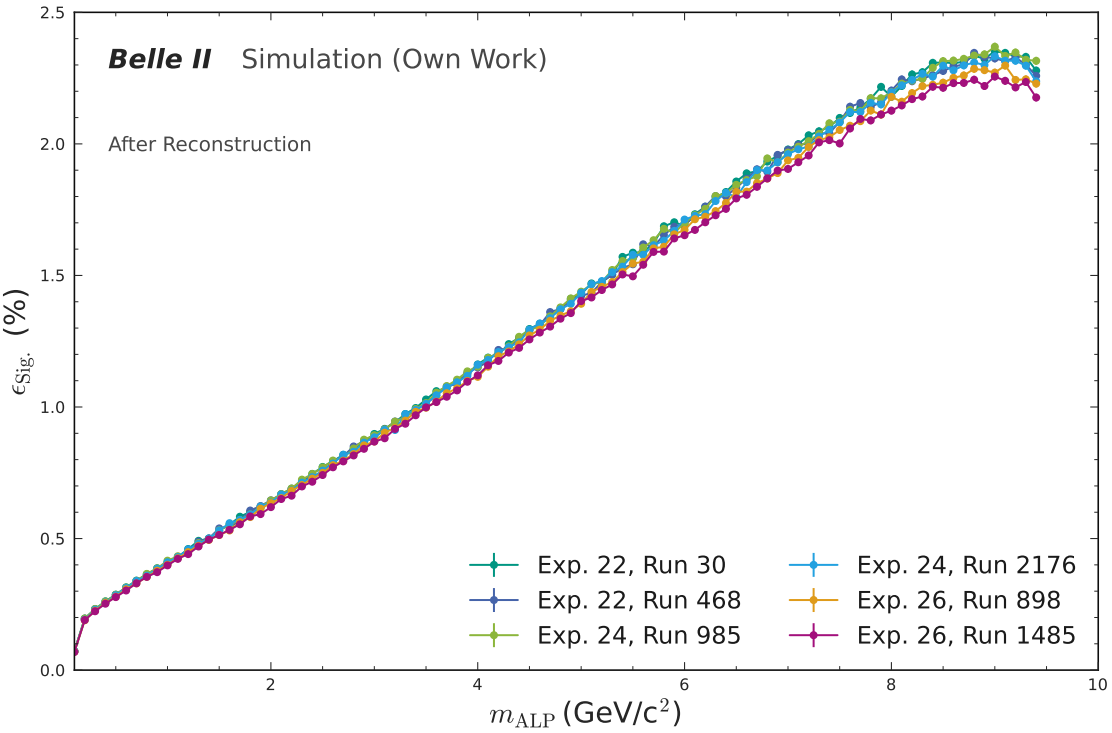


Figure 6.3.: This shows the signal efficiency $\epsilon_{\text{Sig.}} = \frac{n_{\text{Rec.}}}{n_{\text{Gen.}}}$ over the ALP mass m_{ALP} after the reconstruction for the 6 runs of signal MC. The efficiency rises approximately linearly with the ALP mass from $\approx 0.2\,\%$ to $\approx 2.5\,\%$ for all studied runs. The highest and lowest masses, however, deviate from this scaling by showing a steeper decrease in efficiency.

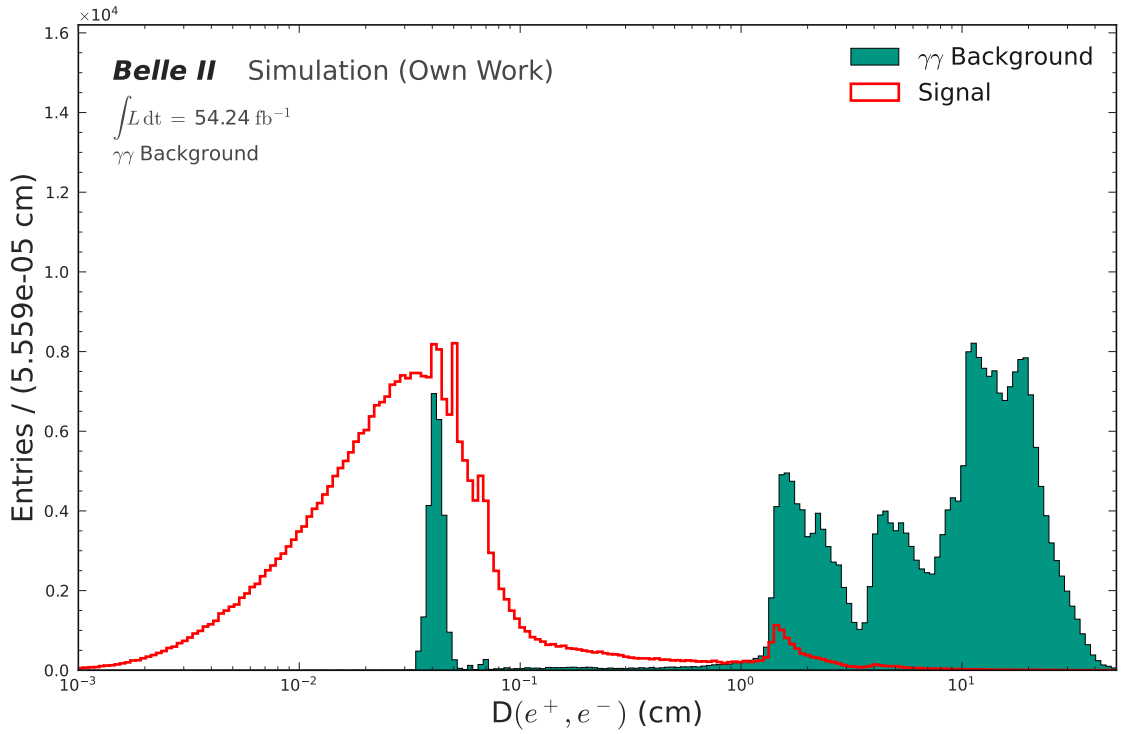


Figure 6.4.: e^+e^- vertex displacement obtained from the TreeFit [31] for the signal and $e^+e^- \to \gamma\gamma(\gamma)$ background MC.

To specifically suppress the background from $e^+e^- \to \gamma\gamma(\gamma)$, which is very similar to the s-channel signal signature, a selection on the vertex displacement is used. As explained above, the signal e^+e^- vertex is expected to be in the interaction point, while the background has a $\gamma \to e^+e^-$ conversion in material, which is displaced. Figure 6.4 shows that the vertex distance $\mathrm{D}(e^+, e^-)$ has the expected behaviour, and most of the signal events feature low displacement distances, while the $\gamma\gamma$ background has larger values. A selection at $\mathrm{D}(e^+, e^-) < 1\,\mathrm{cm}$ removes 85 % of the $\gamma\gamma$ background component, while retaining 95 % of the signal.

Another challenging background component are charged pions and kaons reconstructed as leptons, especially from the $e^+e^- \to q\bar{q}$ backgrounds. To differentiate charged particles, basf2 offers different versions of PID variables that combine information from different subdetectors. This analysis uses an implementation of the global PID based on a Boosted Decision Tree (BDT) to identify whether the reconstructed electrons and positrons are proper electrons and positrons [36]. Figure 6.5 displays the distribution of this variable for the electron $y_{\mathrm{PID}}(e^-)$ for the signal MC as well as the background from the $e^+e^- \to q\bar{q}$ samples. For the background, the plot is split into two parts: one where the reconstructed electron is actually generated as an electron, and one where it is not. It clearly shows that most proper electrons in signal and background are classified with higher values around 1, and fake electrons mostly at 0.

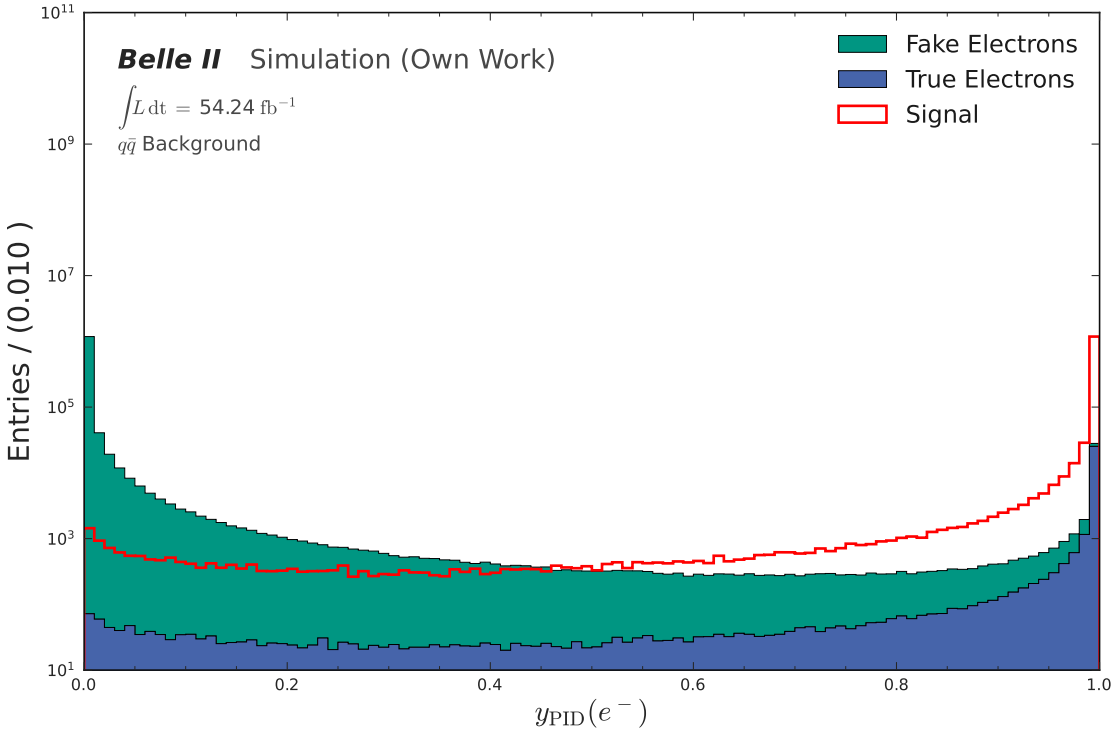


Figure 6.5.: The output score of a decision-tree trained for identification of electrons. While the signal contains only true electrons that mostly get classified with a score of around 1, the $e^+e^- \to q\bar{q}$ background shown contains electrons as well as hadrons, mostly $\pi^+\pi^-$ pairs. The background is split into those groups, and it is evident that the variable classifies most electron candidates correctly.

The best found working point for this selection is the "or" connected one on the electron and positron with a very high threshold value of 0.9. This strategy keeps 99 % of the signal, while removing 95 % of fake leptons, a performance not achieved with other strategies for PID.

As the background level, especially from Bhabha scattering $e^+e^- \to e^+e^-(\gamma)$ after these selections, is still too high to calculate a sensitivity, the next section will present a neural network-based event classifier. This reduces the background to a manageable level, allowing for subsequent sensitivity extraction.

7. Neural Network Event Classifier

This chapter introduces the Neural Network (NN) based event classifier, which targets general background reduction by utilizing a collection of geometric and kinematic variables. Section 7.1 explains the input features and the preparation of the dataset, while section 7.2 covers the setup used for training the classifier. Section 7.3 discusses the performance of the NN classifier and section 7.4 presents the resulting selection strategy from these discussions. Additionally, it examines the resulting background spectrum in the scan variable $M_{\gamma\gamma}$, as well as the signal efficiency and its run dependence, to validate the further extraction strategy in chapter 8.

7.1. Preparation of Dataset

To utilize the full power of the NN as a classifier, the proper input data is crucial for maximal learning performance. This section describes the selected set of variables, methods for calculating the input weights, and other relevant details used to prepare the dataset for the classifier.

The features used to train the network should encapsulate the information necessary to separate the Beyond the Standard Model (BSM) signal events from the Standard Model (SM) background. This requires a tradeoff between increasing the number of variables to maximize information and keeping it low to reduce the input space dimension, as the feature space scales rapidly with the number of features, requiring a much larger dataset to cover the input space sufficiently. The set of 26 features from four different groups, presented in table 7.1, was found to be a sufficient trade-off for this analysis.
First general kinematic features like the energies and momentum (components) of the final-state particles, as well as the reconstructed Axion-Like Particle (ALP), introduce a general separation and help the network to learn the mass scale, as they are correlated with the ALP mass. Ideally, one would use the scan variable $m_{\gamma\gamma}$ for this purpose, but this introduced issues in the training that might be resolvable with very tight sampling of the signal, which would require high amounts of generated signal mass points. Additionally, the similarity between the two photon energies,

$$E_{\text{Simmilarity}} = \frac{E_{\gamma H} - E_{\gamma L}}{E_{\gamma H} + E_{\gamma L}}, \tag{7.1}$$

enters the training, where $E_{\gamma H}$ ($E_{\gamma L}$) for the higher (lower) energetic photon in the event. This variable identifies events where one photon comes from the beam background. In these instances, the beam background photon typically has a low energy, barely passing the $E_\gamma > 0.1\,\text{GeV}$ selection, and is combined with a higher-energy proper photon, leading

to a large asymmetry in energy. In signal, the asymmetry is less pronounced due to the low boost of the ALP.
The second group contains the 3D opening angles $\Phi_{i,j}$ for all combinations of two particles i, j in the event. In the case of track-based particles (e^+ and e^-), the angle definition is ambiguous because the angles change during the trajectory due to the strong magnetic field in the detector. Therefore, these use the reconstructed θ and ϕ angles from the tracking system, which match the ones at the start of their trajectories as well as the angles calculated via extending their helical paths to the point they hit the Electromagnetic Calorimeter (ECL). These two definitions help identify processes that happen along the track, like Bremsstrahlung or pair conversions, better, while also keeping the information of the initial event topology. For the photons, the only meaningful angle is the one calculated from the ECL cluster, as they are neither affected by the magnetic field nor leave tracks in the Central Drift Chamber (CDC). Therefore, the total number of opening angles extends from 6 to 11 as the one between the two photons is only calculated once.
Additionally, the ratio between the energy measured in the calorimeter and the momentum measured by the tracker, E/p, is used as a feature for the electron and positron. E/p helps in Particle Identification (PID) to further remove $q\bar{q}$ backgrounds containing charged pions or kaons and the backgrounds containing muons, which are reconstructed as electrons. While the electrons ideally deposit all their energy in the calorimeter, leading to a ratio of approximately one, other particles interact less in the calorimeter and have a significant amount of energy left afterwards, leading to a ratio below 1.
As a final group, approximations of the Mandelstam variables enter the training. F. Acanfora et al. showed that Mandelstam variables s and t show a strong signal-background-separation between the signal $\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma$ process and radiative Bhabha scattering [8]. In case of neglecting the electron mass, the expressions for those quantities are

$$\begin{aligned} t_\pm &\approx -\sqrt{s}E^*_{e\pm}\left(1 \pm \cos\theta^*_{e\pm}\right), \\ s_\pm &\approx \sqrt{s}\left(\sqrt{s} - 2E^*_{e\pm}\right). \end{aligned} \tag{7.2}$$

Where "*" denotes variables in the collision center of mass frame, and $\sqrt{s}$ is the collision energy [8]. As the collision energy $\sqrt{s}$ is a constant prefactor, it is excluded from the calculations here, and $\pm$ in $t_\pm$ is generalized for a more straightforward calculation. For a given particle $p \in e^+, e^-$, the used formulations are

$$\begin{aligned} t_+(p) &= E^*_p(1 + \cos\theta^*_p) \\ t_-(p) &= E^*_p(1 - \cos\theta^*_p) \\ s(p) &= E_{\mathrm{Beam}} - 2E^*_p. \end{aligned} \tag{7.3}$$

All six of these variables are used as input features for the NN.

As the neural network requires inputs to be in the range between 0 and 1, these features are normalized. To reduce runtime and avoid effects from unphysical missreconstruction, input ranges are manually set rather than calculated automatically from the column-wise minimum and maximum of each variable. With the normalization, the input vector $\vec{X}$ for the network is complete, and the missing components are the event-wise weights and the target labels.

The training utilizes Monte Carlo (MC) events, hence generator-level information is also available. While it makes no sense to use for the input features, as those must also be usable in data, the weights and labels can and should be calculated using MC information. For the labels $\hat{Y}$, which encode whether the network perceives the event as a signal ($\hat{y} = 1$) or background ($\hat{y} = 0$), the provided information from the truth matching is used. While background can be any reconstructed event from background generation, the signal must be entirely truth-matched. Truth-matching means that both reconstructed photons must be generator-level photons, originating from the ALP decay, and that the leptons must be the generated leptons that come immediately from the beam; thus, the reconstructed event corresponds entirely to the generated one. This requirement ensures that the classifier learns the proper signal shape, while the background is as close as possible to actual data.

To equalize potential imbalances between the signal dataset and the background, as well as between the individual masses and background components, input weights are needed. The weights ensure that the proper classification of all components is seen equally during training. For background, the events are first weighted individually to the cross section of the generated process. This again requires the MC information about the process and the number of streams used to generate that subset. The weighting of the different processes ensures that the weighted events the network sees resemble the composition of events expected from the data. On the signal side, the weights are first set to equalize the contribution from each mass, to remove effects from changing efficiency with the mass, and to account for the steps in the number of generated events. In the final step, the signal weight is globally scaled so that the sum of all signal and background weights is equal. This now defines the full weight vector $\vec{W}$, which assigns importance to every input event during training.

However, issues can also arise from a highly imbalanced number of events between signal and background, even if they are weighted to be equally important. Although they have equal importance by the weighting, large imbalances between the classes bring different sampling densities in the feature space, leading to suboptimal classification on larger datasets. To remove potential imbalances, an additional factor for the fraction of signal events that are sampled for training is introduced before the weights are calculated. The sampling allows for approximately balancing the number of signal and background events entering the classifier, while the weights fine-tune it further to achieve exact balance.

Table 7.1.: Input features used for the neural network, and their valid range. The features represent different variables, from straightforward kinematic variables like energy and momenta of the four final-state particles and the ALP to more engineered ones. The ranges are introduced to efficiently normalize the features to the input range of $[0, 1]$, which means some of them extend to unphysical regions to catch outliers and reconstruction errors.

Feature	Range	Unit
E_a	$[0, 20]$	GeV
$p_{T,a}$	$[0, 20]$	GeV/c
$p_{z,a}$	$[-10, 15]$	GeV/c
E_{γ_L}	$[0, 7]$	GeV
E_{γ_H}	$[0, 7]$	GeV
E_{e^+}	$[0, 7]$	GeV
E_{e^-}	$[0, 7]$	GeV
$E_{\text{Simmilarity}}$	$[0, 1]$	1
$\Phi^{\text{ECL}}_{e^+e^-}$	$[0, 3.5]$	rad.
$\Phi^{\text{Track}}_{e^+e^-}$	$[0, 3.5]$	rad.
$\Phi^{\text{ECL}}_{e^+\gamma_H}$	$[0, 3.5]$	rad.
$\Phi^{\text{Track}}_{e^+\gamma_H}$	$[0, 3.5]$	rad.
$\Phi^{\text{ECL}}_{e^+\gamma_L}$	$[0, 3.5]$	rad.
$\Phi^{\text{Track}}_{e^+\gamma_L}$	$[0, 3.5]$	rad.
$\Phi^{\text{ECL}}_{e^-\gamma_H}$	$[0, 3.5]$	rad.
$\Phi^{\text{Track}}_{e^-\gamma_H}$	$[0, 3.5]$	rad.
$\Phi^{\text{ECL}}_{e^-\gamma_L}$	$[0, 3.5]$	rad.
$\Phi^{\text{Track}}_{e^-\gamma_L}$	$[0, 3.5]$	rad.
$\Phi^{\text{ECL}}_{\gamma_L\gamma_H}$	$[0, 3.5]$	rad.
E/p_{e^-}	$[0, 5]$	1
E/p_{e^+}	$[0, 5]$	1
s_{e^-}	$[-2, 12]$	GeV
s_{e^+}	$[-2, 12]$	GeV
$t^+_{e^-}$	$[-1, 11]$	GeV
$t^-_{e^-}$	$[-1, 11]$	GeV
$t^+_{e^+}$	$[-1, 11]$	GeV
$t^-_{e^+}$	$[-1, 11]$	GeV

After these calculations, the signal and background events are mixed and separated into three datasets: training, validation, and test. This approach is common practice to monitor the training more effectively, as the training dataset is only used for training and optimizing the network. The validation dataset is used to calculate the loss and other quality monitoring quantities during the training, but the actual optimization does not use them. In the optimal case, the curves calculated for the loss of the training and validation datasets match. However, if there are issues, the validation dataset typically performs significantly worse than the training set. One prominent example of such an issue is overtraining, where complex classifiers fail to learn the proper decision boundary, instead learning the individual events and their noise when trained for too long. Overtraining results in a loss of generalization to untrained datasets, such as the validation set. Hence, the validation loss rises rapidly, while the training loss decreases. The third dataset, test, is used for monitoring plots after training, but it is not necessary, as these can also be done on the validation dataset.

7.2. Training and Optimization

With the dataset and preparation steps done, the actual classifier can be built and trained. This analysis uses a fully connected feed-forward neural network for this purpose [37, 14]. These work by feeding the input features through a matrix multiplication with the weight matrix and adding a bias vector. The resulting vector is passed to a nonlinear function, known as the activation function, which generates the input vector for the next layer, where the number of nodes corresponds to the length of that vector. This procedure is then repeated to build a new layer several times. During training, these weights and biases are recalculated to improve performance, iterating through the entire dataset in batches multiple times (epochs). For the recalculation, the direction in which the parameters need to change must be calculated by using a loss value that describes the deviation between the predicted output value y and the true labels $\hat{y}$ for each batch. The gradient of the loss value with respect to the weights and biases can be calculated using the chain rule, indicating the direction of optimization. Afterwards, an optimization algorithm calculates the new set of weights that minimizes the loss. If the optimization is repeated multiple times, the prediction converges to the targets, and the loss approaches zero.

However, a successful training is not that straightforward in reality. Beyond the weights and biases, the network has multiple non-trainable parameters, like: number of layers to use, how wide each intermediate layer is, what the non-linear functions are, how large the batches are, how many epochs to train, and several parameters regarding the actual optimizer steps themselves, such as the minimizer algorithm or the learning rate. Those parameters, called hyperparameters, are not known a priori and usually can not be calculated; they have to be found using trial and error. To tune the hyperparameters in an automated fashion, the analysis employs a grid scan, where numerous parameter combinations are trained. However, scanning all parameters would be unfeasible, especially as they are partially correlated and some have prior knowledge of what works and what does not. Especially, the nodes per layer increases the number of parameters to optimize

Table 7.2.: Hyperparameters varied in the scan and their allowed values. The sampling was done using uniform distributions in the allowed range, or the logarithmic range (If stated). The activation functions and the network shapes tested are shown in separate figures 7.1 and 7.2.

Scanned Parameter	Values	Notes
Batch Size	[256, 8192]	Log-spaced
Epochs	[50, 250]	Steps of 10
Width	[36, 256]	Log-spaced
Depth	[6, 15]	
Activation Function	See Plot	
Act. Func. Last Layer	Sigmoid	
Shape	See Plot	

significantly, as there is, in principle, any shape of network possible. Previous applications already showed strong constraints on what works and what does not regarding these layer widths; therefore, the analysis uses predefined shapes of networks shown in figure 7.1 that are entirely classified by four parameters: the depth, the maximum width, the activation function, and the shape's name. As activation functions, the scan uses the ones shown in figure 7.2 for the intermediate layers. These functions show good classifier performance and are popular choices for neural networks. To further limit the scan's parameters, the same function is used for all intermediate layers, and the one for the output layer is always a sigmoid function. The sigmoid function is necessary as the network expects outputs on the interval $[0, 1]$ so they match the dataset labels, which is ensured by the sigmoid function

$$\sigma(x) = \frac{e^x}{1 + e^x}. \tag{7.4}$$

These simplifications significantly reduce the number of parameters for the scan, without losing classifier performance. To monitor those scans, the tool Weights and Biases (WandB) [38] is used. WandB allows for viewing and comparing the different runs using a website, which makes selecting the best-performing run easy.
Initial trials of the scan ranges shown in table 7.2 proved reasonable, leading to multiple runs with good convergence in the scan. From those runs, one with the parameters in table 7.3 is selected because it shows low loss in both training and validation, without obvious training issues, and the resulting output classification appeared desirable. The network resulting from the optimal training is then used to predict the output on the full signal dataset, as well as all available background runs.

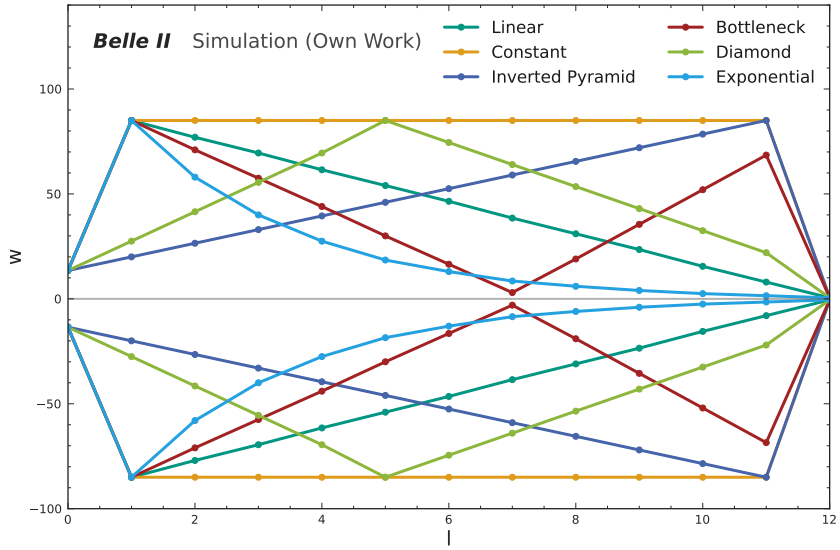


Figure 7.1.: Network shapes used for the classifier sweeps. The parametrization enables the description of the architecture with only five parameters: the number of input features, the activation function, the network width and depth, and the type of shape. This significantly reduces the number of parameters to optimize and allows for finer granularity with low expected impact on the classifier performance.

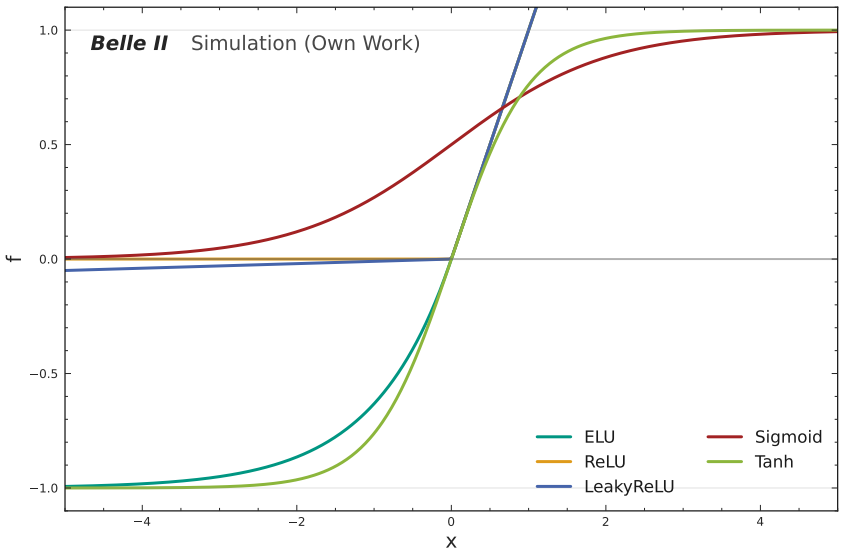


Figure 7.2.: The different activation functions used for the hyperparameter optimization scans. These represent common activation functions used in machine learning applications.

Table 7.3.: Optimized hyperparameters used for the classifier training. These were partially found using parameter sweeps with the values presented in /creftab:scan. Others were kept static throughout the analysis, like the optimizer and scheduler settings, as they are expected to have a minor impact on performance.

Parameter Name	Value
Width	170
Depth	12
Shape	Constant
Activation function	ELU
Epochs	240
Batch size	4608
Optimizer algorithm	ADAM
Learning Rate	10^{-4}
Learning Rate Scheduling	Reduce on Plateau
Loss Function	BCE
Signal Sampling	10 %

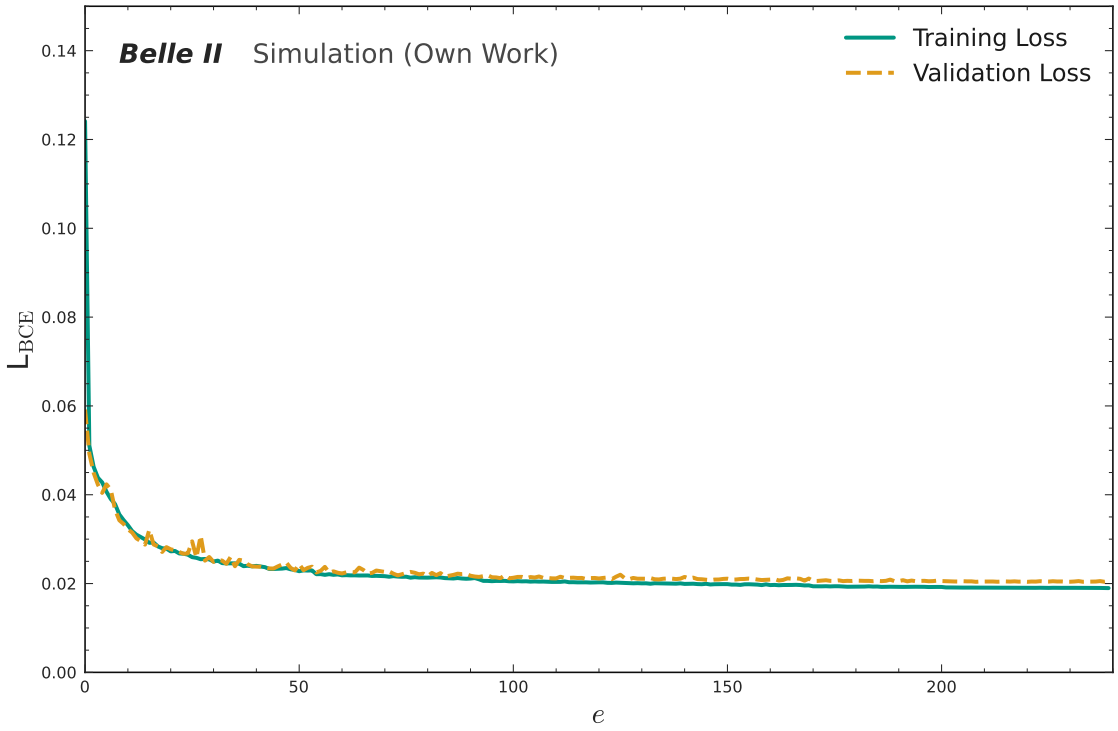


Figure 7.3.: Evolution of the binary cross entropy loss value over the epoch for the setup shown in table 7.3 for the training and validation dataset.
Both show convergence towards a relatively low value quickly, with acceptable spikes in the validation loss. At the convergence, a gap between validation and training loss is visible, but the validation loss is not rising as rapidly as expected for overtraining.

7.3. Classifier Performance

To evaluate the performance of NN classifiers, there are different standard quantities. These should primarily quantify the tradeoff between the false positive and false negative rates of the classification, monitor differences between trained and unseen datasets, and help identify a selection value for the network's output.

A first measure of training quality is the loss on both the training and validation datasets. The value of the loss and the convergence speed can provide hints on whether the classification is sufficient, and the deviation between training and validation can indicate if issues such as overtraining or instability have occurred. This is exceptionally straightforward, as they are conveniently calculated at each epoch e during training, providing insights into the training process over time without the need for additional calculation. Figure 7.3 displays the BCE loss L_{BCE} during the training of the network used for the following sections. The training shows sufficient convergence of training and validation loss towards values lower than in most other setups. The gap between the two curves hints at performance differences between the seen and unseen datasets, but is not further

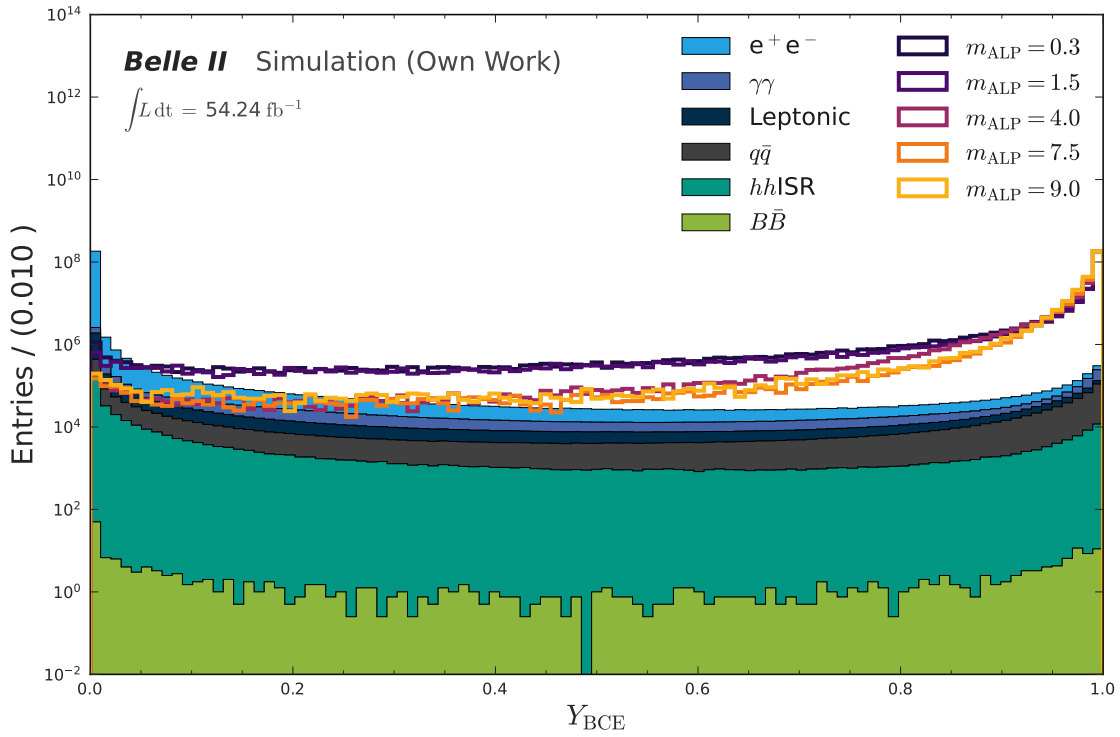


Figure 7.4.: Output value of the neural network classifier for the full available background dataset of 54.24 fb^{-1} and a five signal masses throughout the mass range. Most of the signal (background) events are classified correctly with values close to 1 (0). To better visualize the intermediate regions, the y-axis is logarithmically scaled.

concerning, as it is minimal. Additionally, the validation loss is still decreasing, albeit at a slower rate than the training loss. If severe overtraining occurs, the validation loss would rapidly increase, which does not happen in the current scenario.

To further confirm proper training, the classifier output on the whole available background dataset and exemplary signal masses is shown in figure 7.4. The output indicates that most events are classified correctly, with an output of nearly zero for background or one for signal. However, there are instances of misclassified outputs for both classes, showing at intermediate values and even peaks from signals misclassified entirely to zero predicted output, as well as background to one. Therefore, the selection working point must be derived as a tradeoff between the lost signal and the retained background. To search for a working point, there are different tools available, e.g., the Punzi Figure of Merit (FoM)

$$\mathbf{PFOM} = \frac{\epsilon}{\frac{a}{2} + \sqrt{B}}. \tag{7.5}$$

Where B is the remaining number of background events, ϵ is the signal efficiency, and a is the Gaussian confidence level [39]. The **PFOM** can be calculated for many possible working points along a specific variable, and its maximum value suggests a good tradeoff

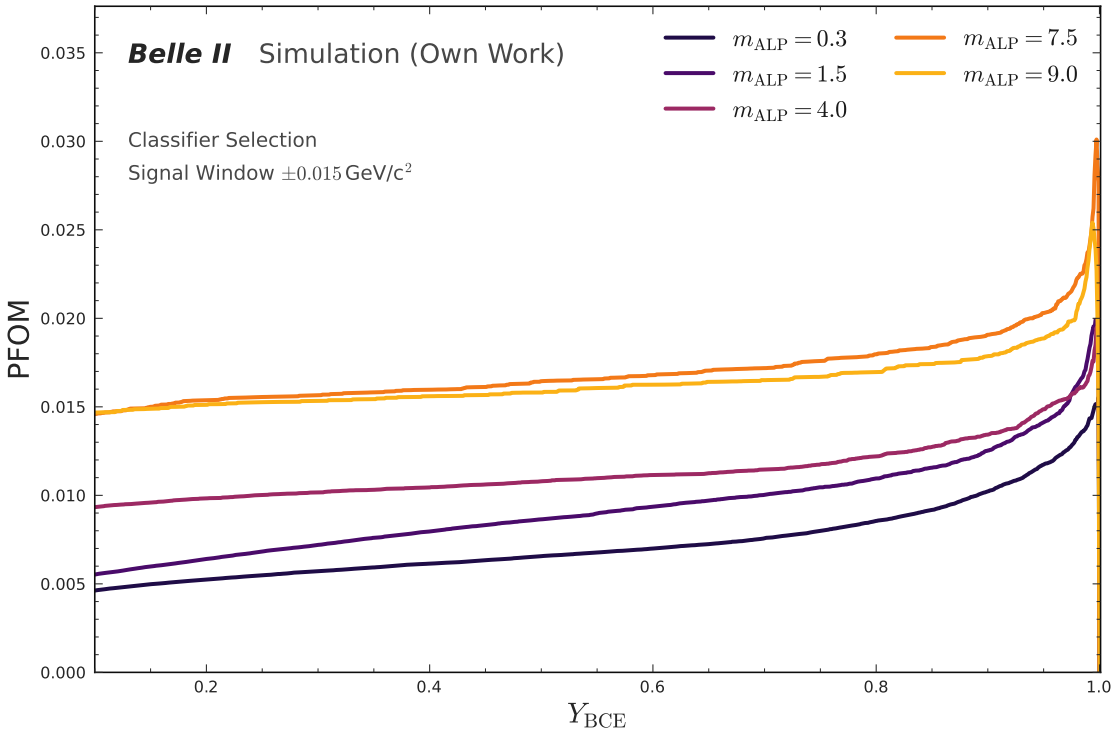


Figure 7.5.: Punzi figure of merit over the neural network event classifier working point evaluated in the $\pm 0.015\,\mathrm{GeV}/c^2$ signal window in $M_{\gamma\gamma}$.

between signal efficiency and remaining background. However, studies of the punzi FoM for the output shown in figure 7.4 suggest high working points very close to 1, as shown in figure 7.5. These are discarded because their performance is highly susceptible to training and may depend on statistical fluctuations between classifiers trained under equal conditions. Ultimately, a flat working point of 0.9 is used throughout the masses, as it contains most of the signal tails.

To clearly visualize the impact of the selection point on the quantities of interest, signal efficiency of the classifier, and relative background rejection, Receiver Operating Characteristic (ROC) curves are standard tools. These usually display the signal efficiency on the x-axis and the background rejection rate on the y-axis for various working points on the classifier, or closely related quantities. As both of the axes are normed to the range of $[0, 1]$, they make measuring the overall performance and visualization of working points straightforward. An ideal classification would yield perfect signal efficiency and background removal, with only losing signal or retaining background at the edge working points of exactly 0 or 1. Hence, such a classifier would result in a rectangular ROC curve. A random classifier, on the other hand, would statistically treat both equal and not favour signal over background, and hence the ROC curve is a straight line. Usually, real curves lie between these two extremes, being a rounded rectangle shape.

One standard measure for the overall classifier power is the Area Under Curve (AUC), which is the integral of the ROC curve. The AUC is bounded between $\frac{1}{2}$, meaning performance as a random classifier, and 1 for the ideal case. The lower bound is not zero, as a

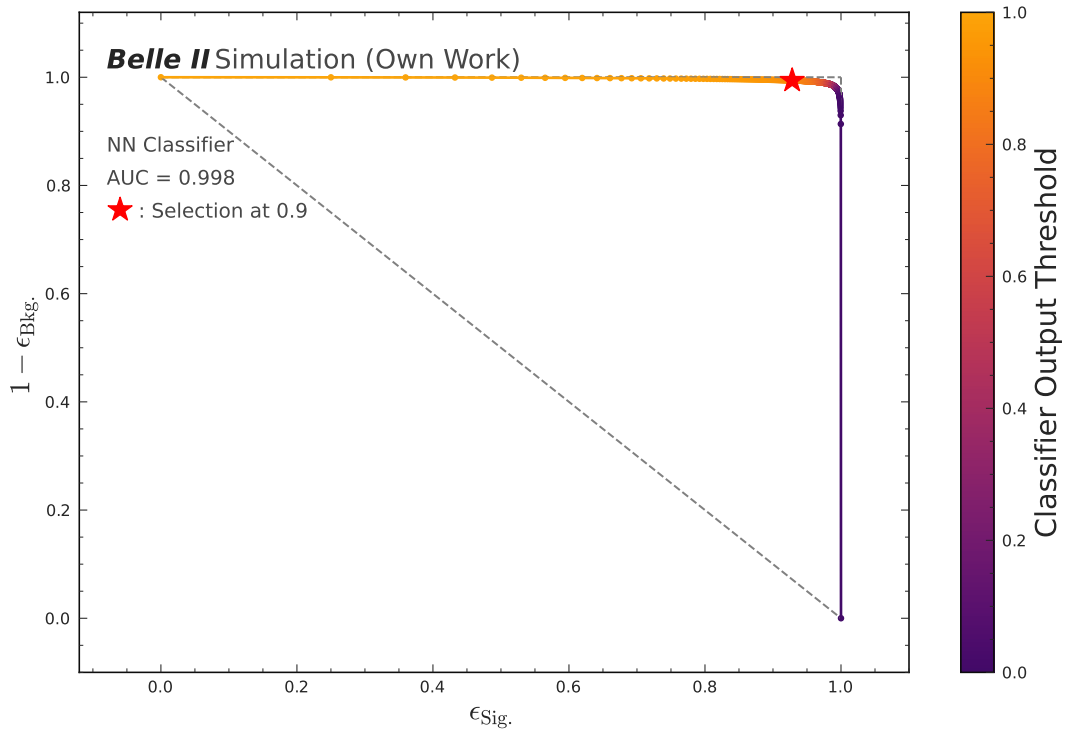


Figure 7.6.: Receiver operating characteristic of the neural network classifier using 1000 equally spaced working points on the intervall $[0, 1]$. Overall, the classifier achieves an area under the curve of 0.998, and the selected threshold of $Y_{\text{BCE}} > 0.9$ is highlighted.

classifier with a performance worse than random could be trivially flipped to achieve a better-than-random performance.
Beyond that, attaching the working point to the curve, e.g., by using a color gradient, can help visualize specific selection points clearer than the output itself. The ROC curve achieved by the used NN classifier in this analysis can be seen in figure 7.6. Its overall performance is exceptionally powerful with **AUC** = 0.998. The highlighted working point for 0.9 is left of the rounded corner, showing that a further reduction of background leads to a rapid decrease in signal efficiency, while an increase of signal efficiency by a looser selection rapidly increases background. However, further attempts to optimize the working point systematically are needed if this analysis progresses.
The next section will use the classifier selection and other selections to extract the signal efficiencies and background distribution in preparation for the sensitivity calculation.

Variable	Condition
Y_{BCE}	> 0.9
$d_{\gamma L}^{\text{c2T}}$	$> 30\,\text{cm}$
$Y_{\gamma_L}^{\text{BB-MVA}}$	> 0.5

Table 7.4.: Signal selection criteria applied after classifier training.

7.4. Signal Selections

With the trained classifier, the selection on the classifier output discussed above can reduce the background almost entirely, while keeping most of the signal events. Additionally, selections targeting the remaining beam background events are applied. Selections achieving a relatively smooth background distribution are beneficial, as they help with the extraction fits in chapter 8.

Table 7.4 shows the selection values used after the classifier training. As discussed, the classifier output selection aims to reduce the general background level. Additionally, remaining events, where the low-energy photon comes from the beam background, are removed by the other selections. To identify those, a central Boosted Decision Tree (BDT) is available that is trained using various cluster shape and timing variables. However, previous analyses observed issues in MC-data-agreement when using the BDT for photons with additional clusters nearby, as the cluster shapes are susceptible to this. To mitigate any issues from nearby clusters, the selection on the distance between the cluster and the closest track $d_{\gamma L}^{\text{c2T}}$ is used, making the decision tree usable for a selection on the predicted score $Y_{\gamma_L}^{\text{BB-MVA}}$. The decision tress output is in the range 0 to 1, where an ooutput score of 1 (0) indicates proper (beam background) photons.

The background distribution in $M_{\gamma\gamma}$ after these selections is in figure 7.7. Compared to the backgrounds after the decay reconstruction, the selections reduce the overall background level by 3 to 4 orders of magnitude. The two most dominant processes in the remaining events are $e^+e^- \to \gamma\gamma(\gamma)$ events not removed via the displacement selection and double radiative Bhabha events. Beyond the performance of the classifier, no clear selection was found to remove the radiative Bhabha events further, without significantly harming the signal efficiency. Especially as the Mandelstamm variables are already strong in suppressing most radiative Bhabha events, as shown by Acanfora et al. [8] or in section 7.4 through the neural network classifier. The events left in the background spectrum after selections are extreme tails of the distributions for those processes and hence mimic the signal very closely, as the initial process of radiative Bhabha scattering is that frequent.

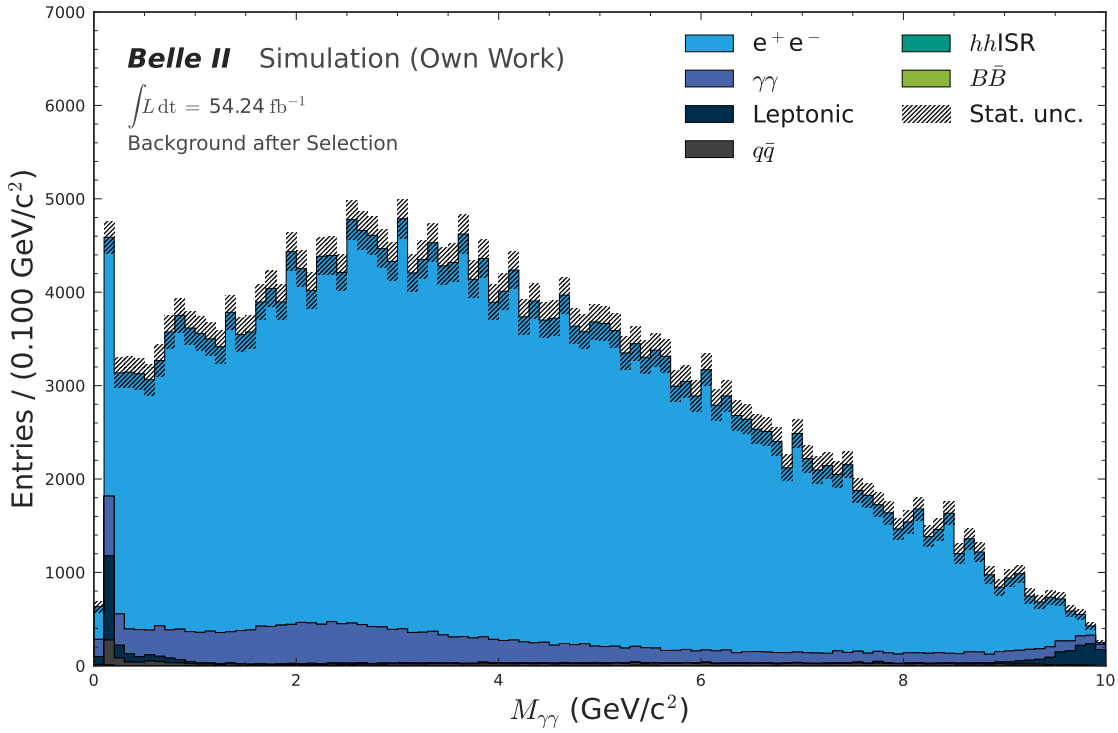
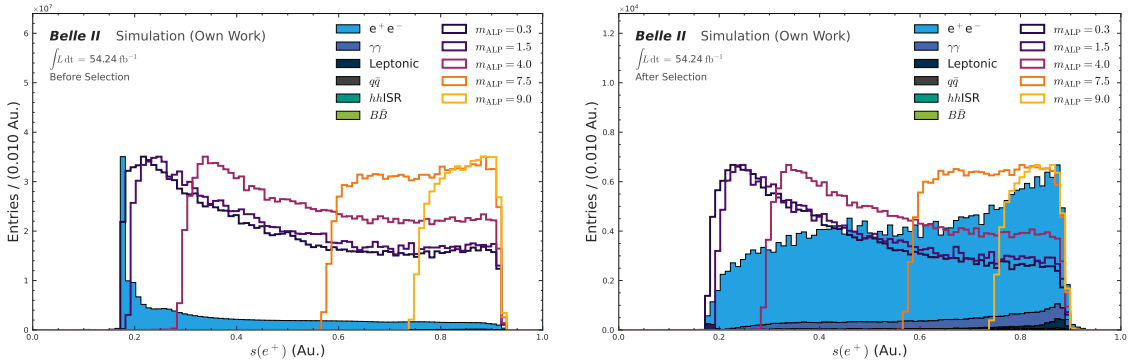


Figure 7.7.: Background distribution in $M_{\gamma\gamma}$ after all selections are applied.



(a) $s(e^+)$ Before selection

(b) $s(e^+)$ After selection

Figure 7.8.: Distribution of the Mandelstam variable $s(e^+)$ for signal and background before and after selection. This is used as an input feature to the neural network, and already normalized to the input range shown in table 7.1 for the plot.

As described previously, the selections aim to minimize signal loss, particularly since the efficiency from geometric constraints is already low (figure 6.3), which limits the final sensitivity. Figure 7.9 displays the signal efficiency

$$\epsilon_{\text{Sig.}} = \frac{n_{\text{Sel.}}}{n_{\text{Gen.}}}, \tag{7.6}$$

after the selection. Here, $n_{\text{Gen.}}$ denotes the number of generated events per mass, and $n_{\text{Sel.}}$ denotes the number of events remaining after selection. The curves show an almost linear rise for the intermediate mass range from about 1 GeV/c^2 ALP-mass up to 7 GeV/c^2. Beyond that, it flattens and decreases again towards the high masses. This behaviour is already present in the efficiency before the selections in figure 6.3, albeit less pronounced. Towards the low masses, a drop in efficiency is also observable.
To better compare how the efficiency changed by the selection, Figure 7.10 shows the proportional deviation of the efficiency after the selections ϵ_{Sel} from the one immediately after decay reconstruction ϵ_{Pre} using

$$\text{L}_{\text{Sig.}} = \frac{\epsilon_{\text{Pre}} - \epsilon_{\text{Sel}}}{\epsilon_{\text{Pre}}}. \tag{7.7}$$

In the intermediate mass region, the loss is relatively flat around 20 to 25 %, while it increases for the low and high masses of up to 50 %.
Noticeably, the selection remains relatively run-independent, as there is no strong difference visible between the run-based efficiencies. To validate the run-independence, figure 7.11 displays the relative efficiency deviation Σ between each run and the averaged efficiency over the runs. There, a small trend is visible leading to lower efficiencies for the higher beam background runs in experiment 26. However the run-dependent fluctuations in efficiency are around ±3 % and hence sufficiently small to ignore for this analysis.

With the three main results from these discussions, the background spectrum in figure 7.7, the signal efficiencies after selection 7.9, and the run independence of it 7.11, the next section will explain the extraction strategy. This allows for the calculation of a mass-dependent expected sensitivity on the cross section, using these three important results.

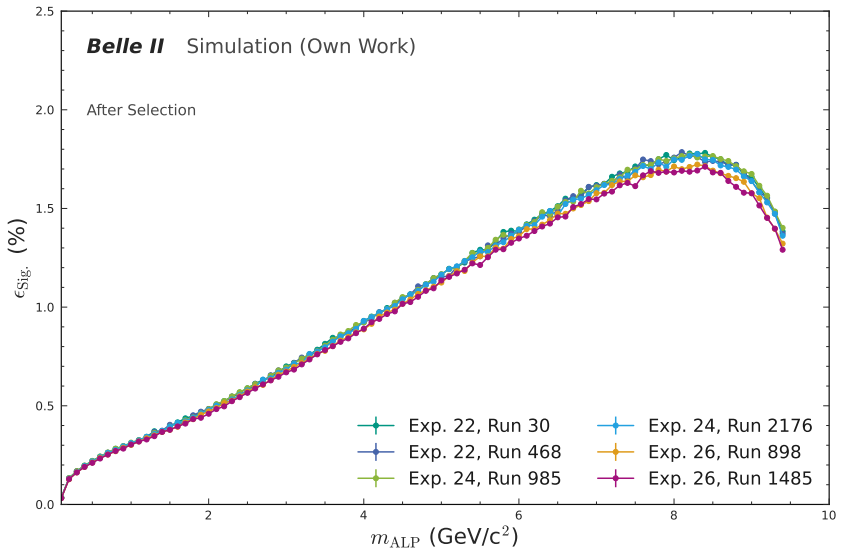


Figure 7.9.: The signal efficiency after selection split by runs over the ALP-mass range.

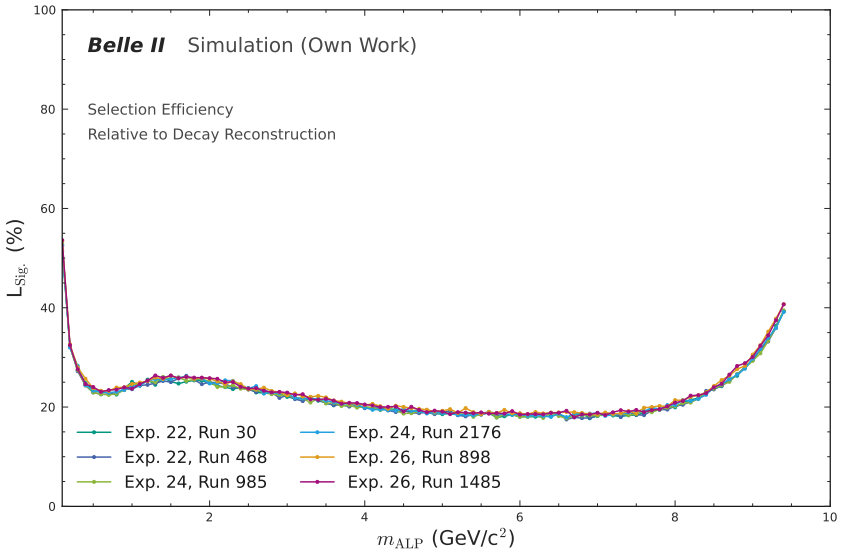


Figure 7.10.: Relative signal loss per mass and run from reconstructed events to the selected ones.

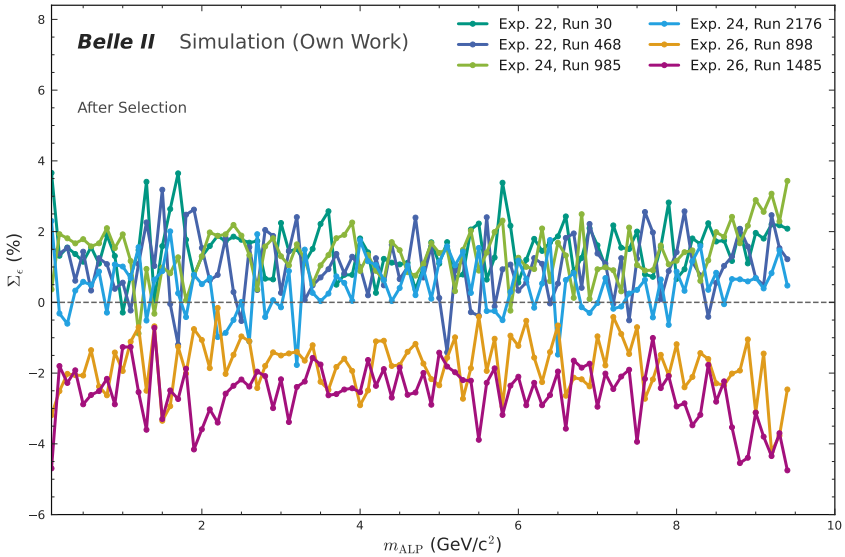


Figure 7.11.: Relative deviation of the run- and mass-wise signal efficiency to the average over the runs for each mass.

8. Sensitivity Extraction

This chapter describes how an expected sensitivity on the cross section is calculated from the selection results discussed previously. The sensitivity fits need the background spectrum, signal width, and signal efficiency as input. First 8.1 shows the signal shape fits after selections, these are subsequently interpolated using the methods described in section 8.2 to evaluate signal shapes and efficiencies on a tighter grid than simulated. Section 8.3 explains the hypothesis fits performed to the Standard Model (SM) background Monte Carlo (MC) with the results from the interpolation and presents the expected sensitivity on the cross section of the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process.

8.1. Signal Shape Fits

To model the signal quantitatively, fits of the signal shape are needed. In particle physics, a common model for modeling signal peaks is the Crystal Ball function, developed by the Crystal Ball collaboration [40]. This function consists of a Gaussian center part with a power-law tail on the left. However, it can also be generalized to a double sided Crystal Ball function which has two independent tails:

$$f(x) = \begin{cases} A_L \cdot \left(B_L - \frac{x-\mu}{\sigma}\right)^{-n} & \text{if } \frac{x-\mu}{\sigma} < -\alpha_L, \\ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) & \text{if } -\alpha_L \leq \frac{x-\mu}{\sigma} \leq \alpha_R, \\ A_R \cdot \left(B_R - \frac{x-\mu}{\sigma}\right)^{-n} & \text{if } \frac{x-\mu}{\sigma} > -\alpha_R, \end{cases} \tag{8.1}$$

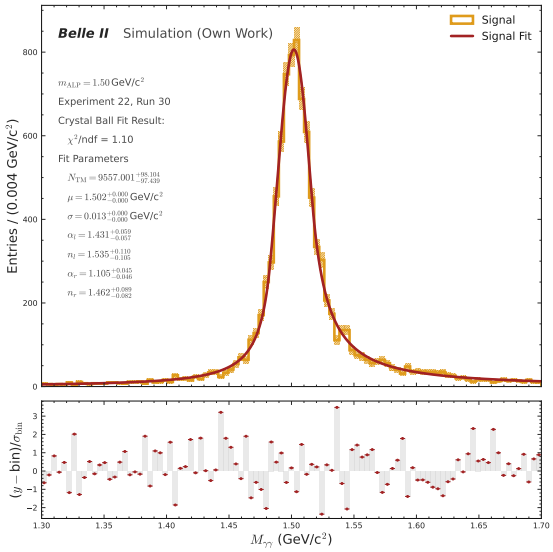
where $A_{L/R}$ and $B_{L/R}$ are defined to ensure contiouous and smooth transition yielding

$$\begin{aligned} A_{L/R} &= \left(\frac{n_{L/R}}{|\alpha_{L/R}|}\right)^{n_{L/R}} \cdot \exp\left(-\frac{|\alpha_{L/R}|^2}{2}\right), \\ B_{L/R} &= \frac{n_{L/R}}{|\alpha_{L/R}|} - |\alpha_{L/R}|. \end{aligned} \tag{8.2}$$

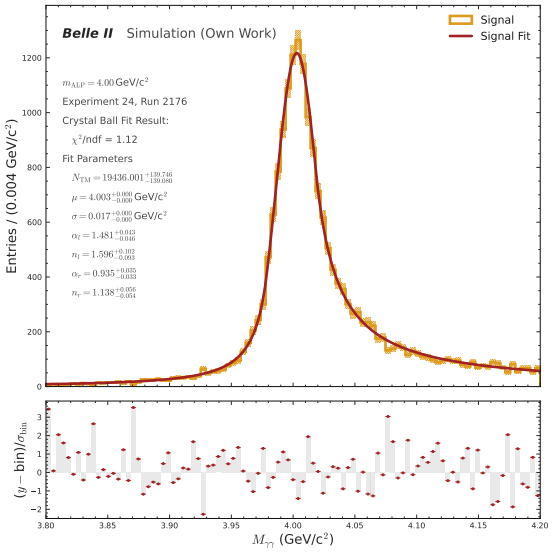
Besides the variable x, which will be the invariant diphoton mass $M_{\gamma\gamma}$, the function has 6 free independant parameters:

- μ and σ describing the central value and width of the Gaussian middle part of the distribution,
- $\alpha_{L/R}$ which determines the transition point to the tails expressed in units of σ away from the central value μ
- $n_{L/R}$ for the steepness of the tails.

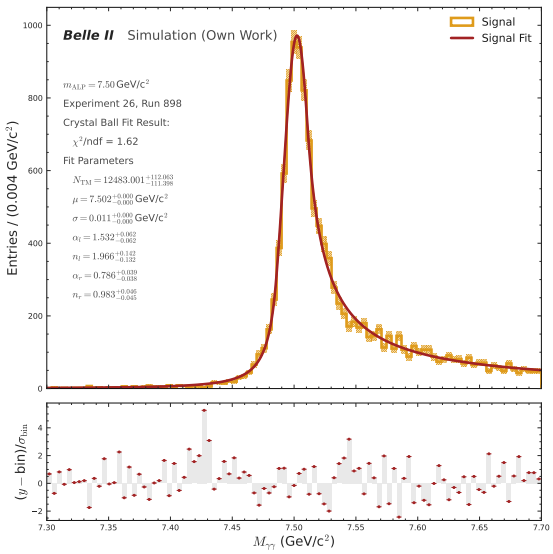
The double sided Crystal Ball function models not only Gaussian peaks well, but also, due to the separate tails, it can also fit sharp peaks with pronounced tails well, as seen in figure 6.1.
The Python package ZFit [41] offers convenient features for fitting in high-energy physics, and has the model for the double sided Crystal Ball function already built in. Figure 8.1 shows the results of some of these unbinned likelihood fits performed to the invariant diphoton mass $M_{\gamma\gamma}$ after selection on the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ signal MC.
Figure 8.2 shows the χ^2 value divided by the number of degrees of freedom for all individual fits. The χ^2/ndf value is a measure of the goodness of the fit and should be close to 1. As the figure shows, this is the case for most of the fits with only a few outliers above 1.5 or below 0.8. Overall, most fits provide a proper description of the signal shapes, and the parameters are presented in the following section, which discusses a challenge due to the small signal width.



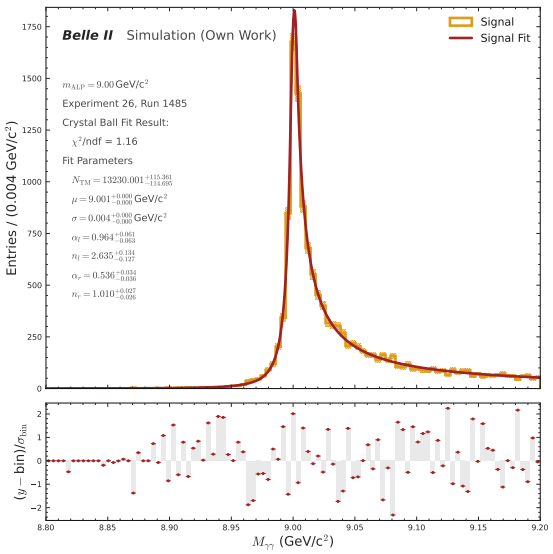
(a) $m_{\text{ALP}} = 1.50$ GeV/c^2, Exp. 22, Run 30



(b) $m_{\text{ALP}} = 4.00$ GeV/c^2, Exp. 24, Run 2176



(c) $m_{\text{ALP}} = 7.50$ GeV/c^2, Exp. 26, Run 898



(d) $m_{\text{ALP}} = 9.00$ GeV/c^2, Exp. 26, Run 1485

Figure 8.1.: Exemplary signal shape fits after selection for different masses and runs. The fits utilize a double sided Crystal Ball function, consisting of a Gaussian-distributed central part with power-law left and right tails, leaving all six fit parameters floating.

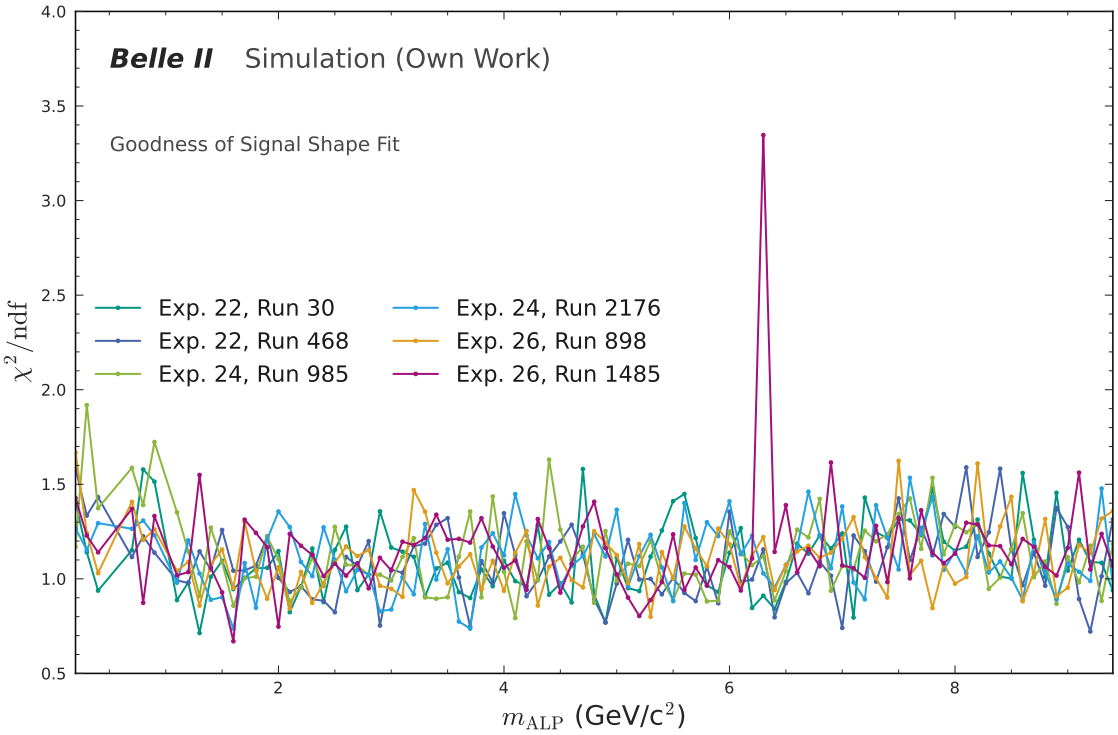


Figure 8.2.: Goodness of fit χ^2/ndf for the signal shape double sided Crystal Ball fits. Most fits are close to the ideal value of $\chi^2/\mathrm{ndf} = 1$ with only a few outliers above 1.5 or below 0.8.

8.2. Interpolation

While the kinematic fit improves energy and mass resolution, which is beneficial for the analysis, it also has a partially adverse effect. Usually, signal MC for a direct search is generated in a way that the spacing between the individual mass points is about the same as the signal width, so the whole scan region is covered. However, due to the fit, the signal becomes narrower ($\sigma_{\mathrm{DSCB}} < 15\,\mathrm{MeV}/c^2$), as shown in figure 8.3b, which would be unfeasible to simulate fully. As the signal efficiency and the fit parameters are considerably smooth throughout the mass range, as shown in figures 7.9 and 8.3, there is a way to mitigate this issue by using interpolation to obtain the signal shape for ungenerated masses.

The interpolation uses fits of all 6 parameters of the double sided Crystal Ball signal shape with Chebyshev polynomials of varying order. These interpolation fits are done using kafe2 [42] as a fitting framework. As the parameters do not have a shape that has a known mathematical description, Chebyshev polynomials are used as an orthogonal basis, which should describe the curves sufficiently well if an appropriate order polynomial is used. The resulting curves from these fits are shown in figure 8.3, and especially the signal width σ is modeled sufficiently well.

These fits can now be used to extract an expected signal shape for any given mass, independent of whether MC is generated for it or not. The generated masses can be further used to calculate the signal efficiency, which is needed as input for the hypothesis test, using the interpolated shapes. From evaluating the interpolated central value μ and width σ at the desired mass value, the signal window can be calculated as $\mu \pm N_{L/R} \cdot \sigma$. Where $N_{L/R}$ has the same values as the hypothesis fit explained section. By selecting signal events remaining after selection within this window, the signal width is calculated. Figure 8.4 shows the curve resulting from the $N_L = N_R = 30$ fit window from the previously discussed signal selection and shape fits. These points are similar to those shown for the full signal range in figure 7.9 and again demonstrate little deviation between the different experiment runs. Assuming the run-dependency of the signal efficiency is negligible, a single curve is fitted to these points as shown in figure 8.4. The Chebyshev polynomial of order 8 describes the efficiency sufficiently well. If one can not assume negligible run dependency, one must either generate a signal MC for each run weighted by luminosity, which is unfeasible, or perform an additional fit over a variable that is suitable for describing the beam background level. For example, the additional hits in the Central Drift Chamber (CDC) or the number of Electromagnetic Calorimeter (ECL) crystals, which measured signals outside of the collision time [26]. Then, each of the shape interpolation and signal fits would be run independently, and a final fit would interpolate the efficiency parameters over the beam background level.

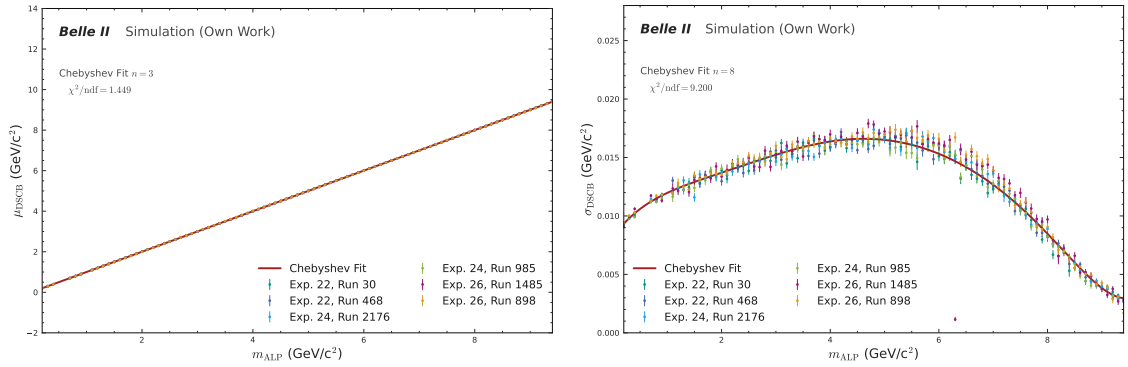
From these interpolation steps, the signal shape and efficiency can be calculated throughout the mass range. However, as the polynomials of arbitrary order have no meaning outside the generated range, they can not be used to extend the reach but only to reduce the granularity of the hypothesis scan. The following section will apply these results to calculate the expected sensitivity on a tight grid and derive expected upper limits on the $e^+ e^- \rightarrow e^+ e^- a,\ a \rightarrow \gamma\gamma$ cross section.

8.3. Hypothesis Fit

While the previous results enable a parametric description of the signal shape and efficiency, the background is not yet defined using a function. To extract the expected background level and shape in the vicinity of the signal resonance for any given mass hypothesis, the background shape must be fitted as well before the hypothesis test can be performed. These fits are performed for all masses on the tight grid achievable through the interpolation, independently. As the fit should be independent of whether a signal of that mass is present in data, the region is first split into a wider fit region, e.g., $\pm 30\sigma$, which is also used to calculate the signal efficiency in the previous section, and an inner signal region of $\pm 3\sigma$ containing most of the signal peak. Events outside the fit region are discarded for the fit, as they do not contribute to the given mass hypothesis. The first fit determines the shape of the SM background; therefore, the inner mass range should not contribute The blinded region ensures that the background shape is not affected by the presence of a signal in the data. As the background shape has no known mathematical description, the fit also uses Chebyshev polynomials as an orthogonal basis to describe an arbitrary function. Order 3 polynomials perform well to approximate the shape. The background function is subsequently fixed, and the inner region is unblinded for the next fit. As this fit now includes a potential signal, the Probability Density Function (PDF) must account for that. Hence, the second fit contains the signal shape obtained via interpolation in the previous section, with fixed parameters, leaving only the signal strength floating. This fit accounts for signal-like fluctuations in the background that impact the final sensitivity. Figure 8.5 shows the results of some of these background fits. While most masses show a low signal-like component as expected, some like 1.5 or $4.0\,\mathrm{GeV}/c^2$ show non-zero fit results in the background due to fluctuations in the inner region.

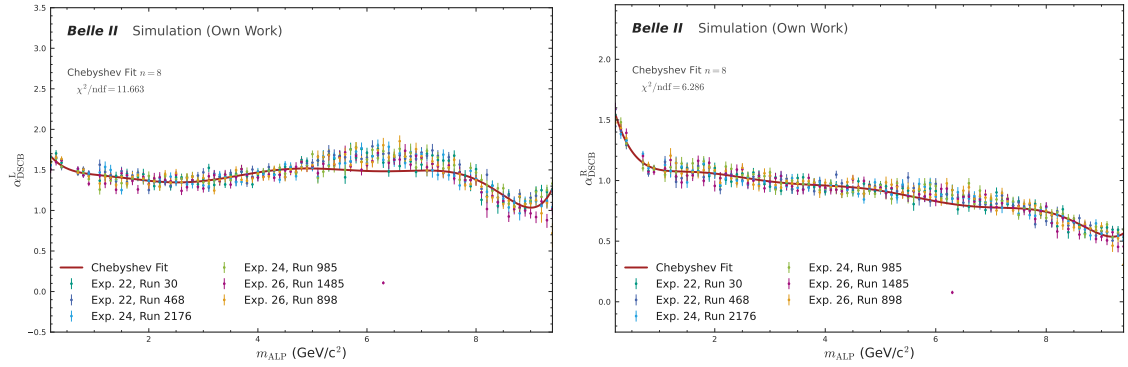
To calculate the test statistics in the presence of a small signal, the first approach is to inject signal MC and recalculate the fits. However, this previously proved to be very costly, hence this analysis uses the asymptotic calculations derived by Cowan et al. [43]. These allow for calculating limits without toy datasets and are therefore more performant. The detailed derivation of the formulae, as well as the assumptions under which the approximation is valid, are given in the publication. As the asymptotic method is fairly common in particle physics, the Python package hepstats [44] offers a convenient and proven implementation that this analysis uses.

Figure 8.6 shows the result of these hypothesis fits, expressed as 90 % upper limit on the cross section for the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process over the mass, together with the expected $\pm 1\sigma$ and $\pm 2\sigma$ uncertainty bands. The cross section limit decreases for higher masses, which is expected as the background decreases, the signal becomes narrower, and the signal efficiency rises. Three factors that benefit a stronger sensitivity on the cross section.



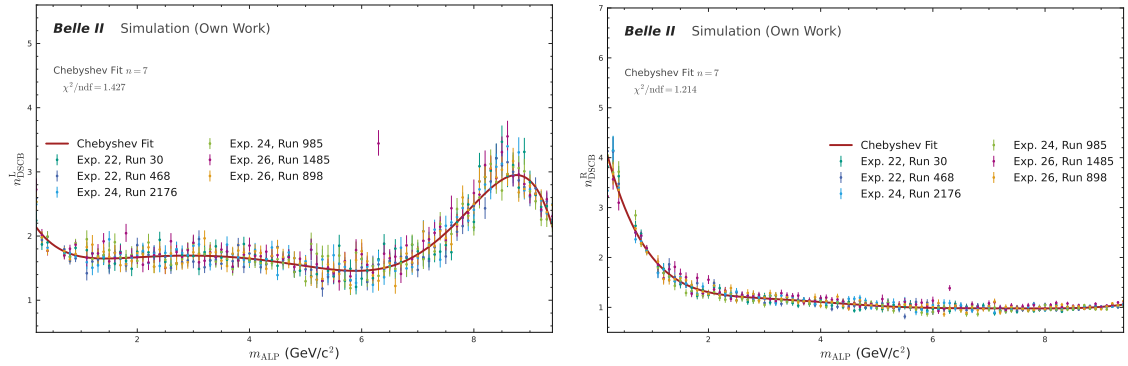
(a) μ

(b) σ



(c) α_L

(d) α_R



(e) n_L

(f) n_L

Figure 8.3.: Chebyshev fits of the double sided Crystal Ball signal shape parameters over the ALP-mass.

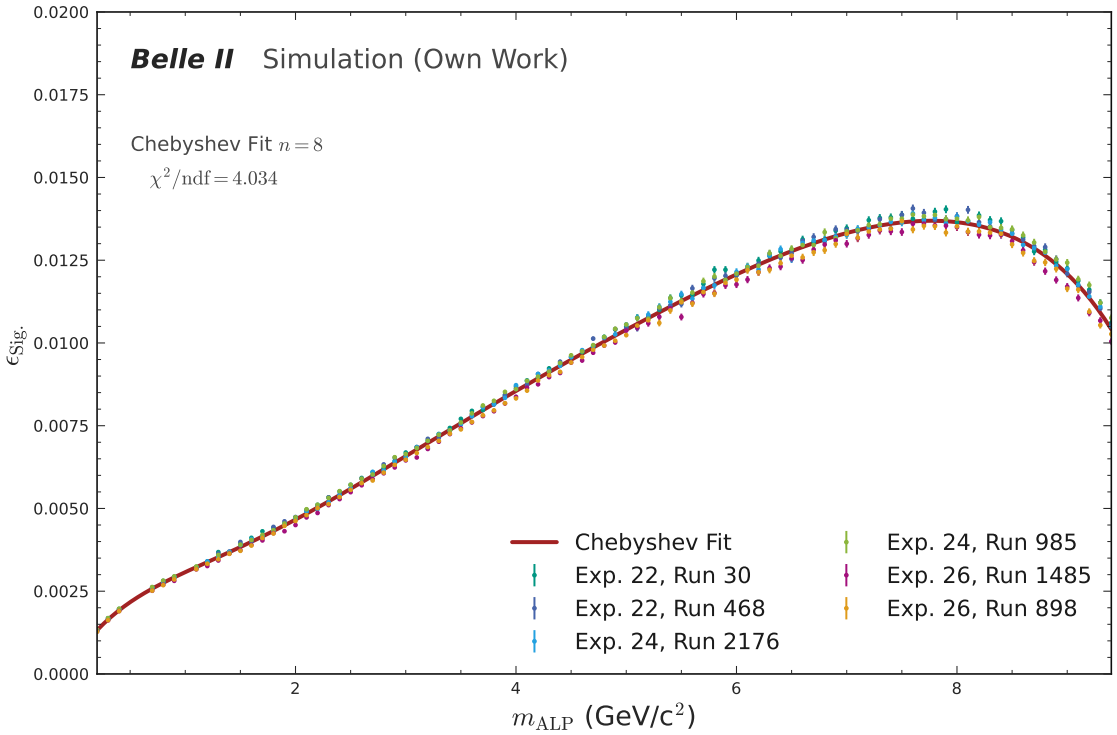
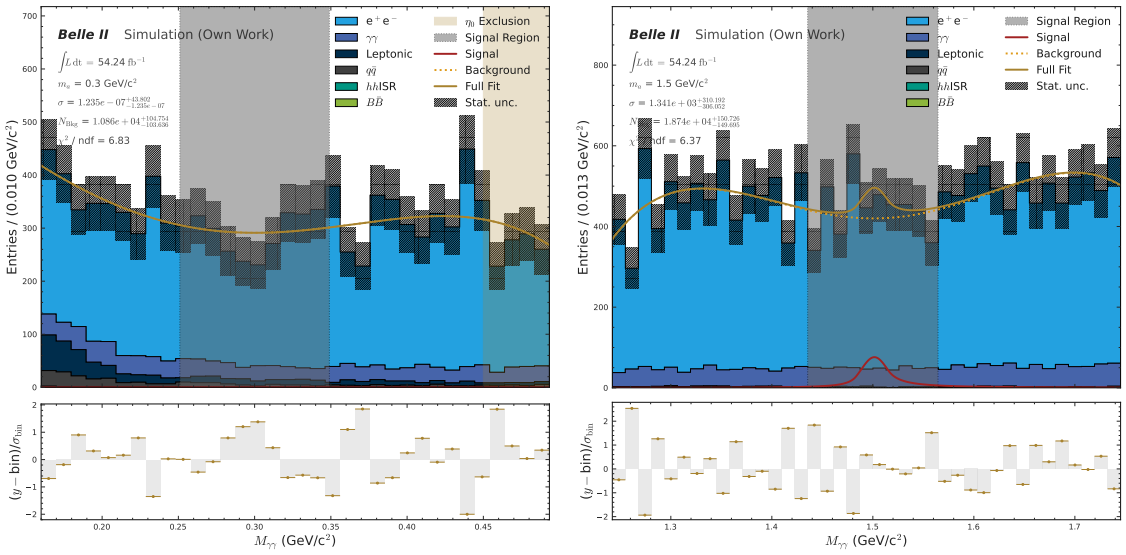
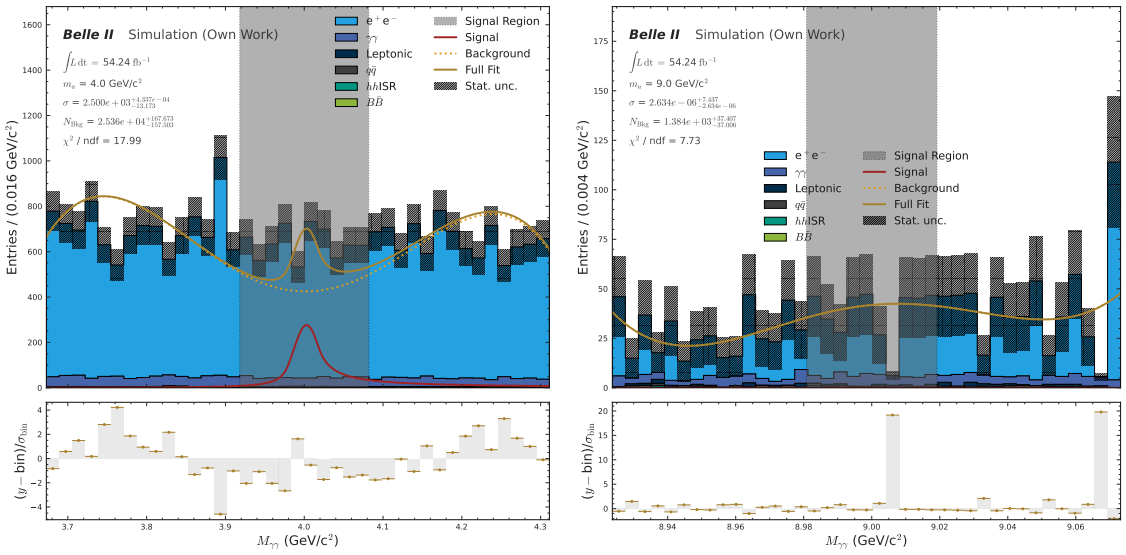


Figure 8.4.: Chebyshev fit to the signal efficiency after selection in the $\pm 30\sigma$ signal window.



(a) $m_{\text{ALP}} = 0.3\,\text{GeV}/c^2$ (b) $m_{\text{ALP}} = 1.5\,\text{GeV}/c^2$



(c) $m_{\text{ALP}} = 4.0\,\text{GeV}/c^2$ (d) $m_{\text{ALP}} = 9.0\,\text{GeV}/c^2$

Figure 8.5.: Fits of a combined background (Chebyshev order 3) and signal (Double sided Crystal Ball) function to the standard model background-only Monte Carlo for four exemplary masses.

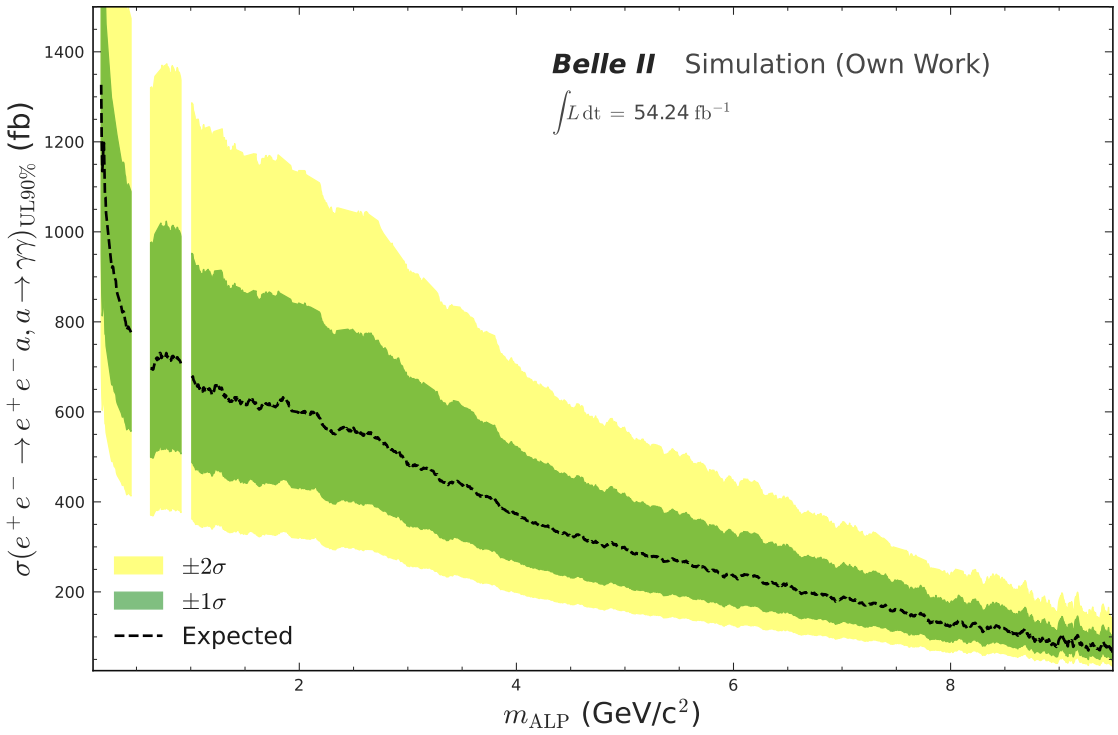


Figure 8.6.: Expected 90 % upper limit on the cross section σ in femto barn for the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process using a dataset of $\int \mathcal{L} dt = 54.24\,\mathrm{fb}^{-1}$, with the $\pm 1\sigma$ ($\pm 2\sigma$) uncertainty region as a green (yellow) band.

9. Model-Dependent Interpretation and Outlook

After the previous sections described the underlying theory model, the experimental setup, and especially the analysis strategy, this section will focus on a model-dependent interpretation of the obtained results in section 9.1. Afterwards, section 9.2 gives an explanation of further possibilities arising from the final state signature $e^+e^-\gamma\gamma$, an extension of this search, and the measurement of standard model contributions.

9.1. Model-Dependent Interpretation

To better compare the sensitivity on the cross section obtained for the $\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma$ channel, shown in figure 8.6, a model-dependent interpretation can convert it into the sensitivity on the Axion-Like Particle (ALP)-photon-coupling. The conversion into a sensitivity $g_{a\gamma\gamma}$ allows to display the result overlayed with the exclusion plot shown and explained in section 2.4.
As the relation between cross section σ and ALP-photon-coupling $g_{a\gamma\gamma}$ is not known analytically for the photon fusion diagram. However, the cross section scales with the square of the coupling as the diagram contains two ALP photon vertices. Hence,

$$\sigma(\mathrm{e}^+\,\mathrm{e}^- \to \mathrm{e}^+\,\mathrm{e}^-\,a,\ a \to \gamma\gamma) \propto g_{a\gamma\gamma}^2. \tag{9.1}$$

This can be trivially expressed as the ratio being a mass-dependent constant. Therefore, the experimental upper limits on the coupling $g_{a\gamma\gamma}^{\text{Exp}}$ are calculable if the experimental cross section limit $\sigma_{\text{UL}}^{\text{Exp.}}$, as well as the theoretical coupling $g_{a\gamma\gamma}^{\text{Theo.}}$, which is a generator setting, and the theoretical cross section $\sigma^{\text{Theo.}}$, which is numerically calculated by Madgraph [1] during event generation, are known via

$$\begin{aligned} \frac{\sigma_{\text{UL}}^{\text{Exp.}}}{\sigma^{\text{Theo.}}} &= \frac{\left(g_{a\gamma\gamma}^{\text{Exp}}\right)^2}{\left(g_{a\gamma\gamma}^{\text{Theo.}}\right)^2}, \\ \Leftrightarrow g_{a\gamma\gamma}^{\text{Exp}} &= g_{a\gamma\gamma}^{\text{Theo.}} \sqrt{\frac{\sigma_{\text{UL}}^{\text{Exp.}}}{\sigma^{\text{Theo.}}}}. \end{aligned} \tag{9.2}$$

Besides the experimental limit on the cross section, the only other non-trivial input is the theory cross section $\sigma^{\text{Theo.}}$, as this is not available on the tight mass grid used via interpolation, but only for the generated masses. A small-scale generation with $n = 10^4$ events is performed on the same $5\,\text{MeV}/c^2$ spacing used in the sensitivity fits to

calculate the theory cross section. The number of events is required as it improves the uncertainty of the calculation and, consequently, yields better results. With 10^4 events, the theory uncertainty is below 0.5 % for all masses, and hence negligible. The calculation with MadGraph5 is necessary, as the theory cross section shown in figure A.1 proved challenging to fit using a single function, and simple approaches like spline-interpolation are more challenging for uncertainty propagation. However, these calculations, given the limited number of events, are sufficiently fast. Additionally, by removing the events from the Les Hoches Events (LHE) output files and keeping only the header with the settings and cross section result, the file size is nearly zero.

With the cross section calculated, the input quantities of equation (9.2) are all known, and the expected experimental sensitivity can be calculated Figure 9.1 shows the obtained sensitivity on $g_{a\gamma\gamma}$ overlayed with the previous analyses. To estimate the current reach of Belle II dataset, the result is upscaled from the $54\,\mathrm{fb}^{-1}$ background dataset to the currently recorded luminosity of $569\,\mathrm{fb}^{-1}$ using the scaling of the coupling with the $\sqrt[4]{\int \mathcal{L}\mathrm{d}t}$. The sensitivity reaches previously unexcluded regions, exceeding the result of the earlier Belle II analysis on ALP-Strahlung $e^+\,e^- \to \gamma\,a,\ a \to \gamma\gamma$ [13], and Beijing Spectrometer III (BES III) [15] throughout the masses. This proves that the analysis of the fully resolved $e^+\,e^- \to e^+\,e^-\,a,\ a \to \gamma\gamma$ process at Belle II is possible.

However, when compared to the sensitivity of the ongoing reanalysis of the ALP-strahlung $e^+\,e^- \to \gamma\,a,\ a \to \gamma\gamma$ process, shown in grey [14], and scaled to the same luminosity, the $e^+\,e^- \to e^+\,e^-\,a,\ a \to \gamma\gamma$ channel clearly falls behind in sensitivity for most masses. This is primarily due to the low signal efficiency for the t-channel process originating from the geometric constraints of the Belle II track acceptance.

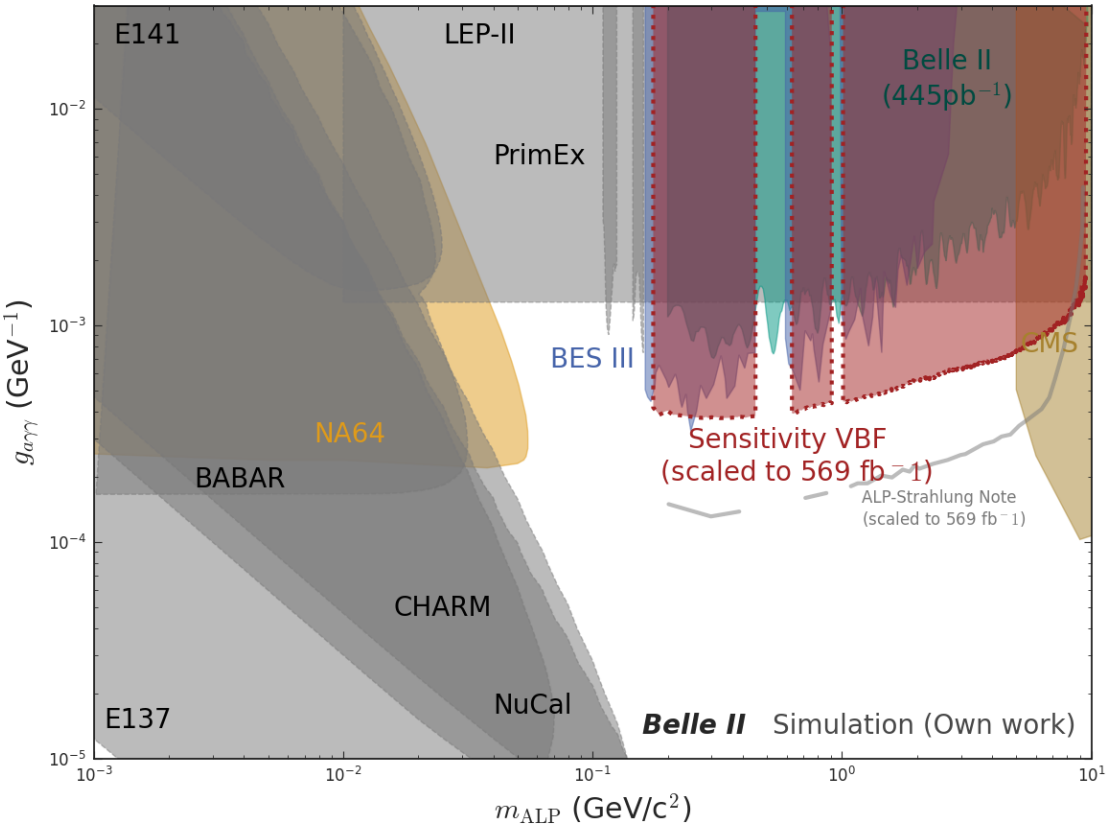


Figure 9.1.: Expected Sensitivity on ALP photon coupling from $e^+ e^- \to e^+ e^- a$, $a \to \gamma\gamma$ on the current Belle II dataset, overlayed with previous analyses and the ongoing re-analysis of ALP-strahlung [14].

9.2. Further Possibilities for the $e^+e^- \longrightarrow e^+e^-\gamma\gamma$ Signature at Belle II

This analysis focused on demonstrating the feasibility of searching for ALPs in the fully reconstructed $e^+ e^- \longrightarrow e^+ e^- a,\ a \longrightarrow \gamma\gamma$ channel. However, it can be extended further by including partially reconstructed states.
The extension is especially promising as the signal acceptance, which is limited by the Belle II track acceptance in θ, is the central limit for the final signal efficiency and hence sensitivity. Extending the search to the final states where only one or no lepton is within the track acceptance would significantly increase the signal acceptance, and hence promise stronger sensitivity, especially as sensitivity usually scales with the square root of the efficiency, while background and luminosity are only entering in the fourth root.
Figure 9.2 shows the acceptance over the ALP masses for all four different reconstruction cases using generator-level information. It clearly shows that even the cases with a single measured electron (positron), called e^- (e^+) tag exceed the acceptance of the fully reconstructed two tag case (0.2 - 2 %) greatly with a range from $1 - 10$ % each. The zero tag case, in which only the photons are in the acceptance, is even better here with 8 to 40 %.

The extension, however, is non-trivial and brings additional challenges to the analysis. On the signal side, the final state is no longer fully reconstructed. Therefore, the information about the missed particles can not be used to separate the signal from the background. In the single-tag cases, the missing momentum vector helps, but it has less resolution than the measured particles. In the zero-tag, the missing momentum provides information only about the combined electron-positron system. Additionally, the overall resolution is severely degraded, as the ability to perform a kinematic fit, which significantly boosts resolution, e.g., in the scan variable $M_{\gamma\gamma}$, as shown in figure 6.1, is no longer available.
Beyond the limitations of the not fully reconstructed final states, the background dataset poses further challenges. The centrally produced background dataset applies different pre-selections in the generation stage to keep the file size bounded and not generate unnecessary undetectable events. One of these selections is the requirement of an electron and positron θ angle above 10° in the Bhabha background generation. The angular selection is further necessary, as the used generators are not validated in the region below, and the differential cross section rapidly diverges. Hence, the analysis would have to generate private background samples, posing not only significant computing efforts, but also these need to be validated against data to ensure proper results of the analysis.
While possible, these extensions exceed the scope of this work, and it is entirely unclear what sensitivity they could achieve.

Given that the semi-reconstructed channels achieve comparable sensitivities to the fully-resolved one and ALP-Strahlung, the searches could be combined. In the absence of a significant excess, this would strengthen the limit and achieve a stronger exclusion than the individual processes. In the event of observation of an excess, these channels could independently confirm the observation.

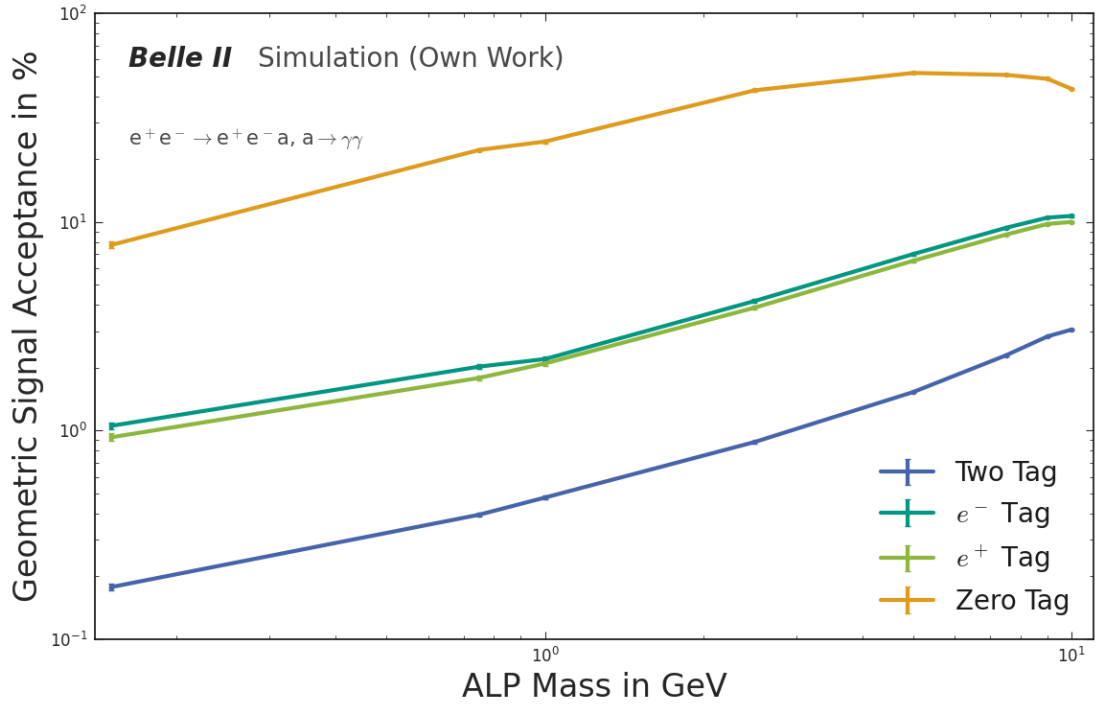


Figure 9.2.: Expected geometric signal acceptance for different reconstruction cases of $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ at Belle II, calculated from generator level information on the final state lepton and photon angles.

10. Conclusion

The thesis at hand presents the first full experimental sensitivity study for Axion-Like Particles (ALPs) in the photon-fusion $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ process at Belle II. Investigating the photon-fusion channel is motivated by a tension between two phenomenology papers: one states that the search is unfeasible [7], and another finds that it yields results comparable to or better sensitivity than the similar $e^+ e^- \to \gamma a,\ a \to \gamma\gamma$ process [8]. While both publications present phenomenological results using limited backgrounds and experimental simulations, this is the first approach to study the $e^+ e^- \to e^+ e^- a,\ a \to \gamma\gamma$ signature using full experimental reconstruction and Standard Model (SM) backgrounds. In contrast to the ALP-strahlung channel $e^+ e^- \to \gamma a,\ a \to \gamma\gamma$, on which Belle II published an early analysis [13], photon-fusion is majorly limited by a low signal efficiency. The limited efficiency primarily results from the final state lepton angles, as the t-channel process favors low scattering angles and hence the leptons are outside the Belle II track acceptance.

The studies show notable SM background contributions from four processes: radiative Bhabha scattering $e^+e^- \to e^+e^-(\gamma)$, diphoton processes $e^+e^- \to \gamma\gamma(\gamma)$, fake leptons from charged pions in quark production $e^+e^- \to q\bar{q}$ and beam background photons. The latter three processes can be suppressed by exploiting their physics using simple cut-based selections. Diphoton processes, for example, must undergo a physical pair conversion in material to produce the e^+e^- pair. Hence, the vertex of the e^+e^- pair is displaced, and a selection on the distance between the vertex and the interaction point removes most of these events. The Bhabha events, which are the majority of the background events, are not easily suppressible using mass-independent selections. A Neural Network (NN) event classifier targets those events by using kinematic and trigonometric variables together with approximations of Mandelstam variables.
This approach reduces the SM backgrounds to a level that makes sensitivity estimation possible, without significantly lowering the geometrically limited signal efficiency further. To calculate the expected sensitivity, the analysis uses a fit-based approach, yielding results in the same order of magnitude as those obtained for the $e^+ e^- \to \gamma a,\ a \to \gamma\gamma$ search.

As stated before, the sensitivity is primarily limited by the low signal efficiency and could be improved by considering partially reconstructed final states in the future. These have a much higher geometric acceptance of the signal and hence potentially higher signal efficiencies. However, the background is not readily available due to selections in the central Monte Carlo (MC) production and hence must be privately generated using dedicated low-angle generators. This procedure is time-consuming as the generator setup must be thoroughly validated to yield results in agreement with data and the other generators used centrally. Therefore, this analysis only investigated the fully reconstructed final state.

In the fully reconstructed case, the result exceeded expectations from [7], demonstrating that an analysis of $e^+ e^- \to e^+ e^- a$, $a \to \gamma\gamma$ at Belle II is possible and feasible. The sensitivity opens the possibility of using photon fusion and ALP-strahlung as cross-checks if one of the analyses finds a significant excess.

A. Appendix

A.1. Signal MC Generation

During the setup of this analysis, two unexpected behaviours in the `MadGraph5_aMC@NLO` generator (Version 3.4.0) were found. Firstly, the cross-section calculation yields unphysical results when used with the Initial State Radiation (ISR) correction and improper generator settings. The second one is a hard limit of 10^6 events per call on MadGraph5.

A.1.1. ISR Correction and Cross Section

The used `MadGraph5_aMC@NLO` generator (Version 3.4.0) Monte Carlo (MC) generator allows for calculating the effects of ISR. This changes the resulting event kinematics, but without adding the ISR photon as MC particle to the event. This is not problematic as the photons are mostly outside of the θ acceptance and additionally very low energetic. Noteably the reanalysis for Axion-Like Particle (ALP)-Strahlung investigated this deeper and found that the higher energetic photons ($E_\gamma > 0.1\,\mathrm{GeV}$) in the ISR spectrum are almost always at extreme θ angles of $\approx 0°$ or $\approx 180°$, hence far outside the accceptance.[14] So the effect of the missing photons is negligible.
However, when enabling the correction, the cross section calculated by MadGraph5 for the photon fusion $\mathrm{e^+\,e^-} \to \mathrm{e^+\,e^-}\,a,\ a \to \gamma\gamma$ process started to show a discontinuity at $m_{\mathrm{ALP}} \approx 2\,\mathrm{GeV}/c^2$ as shown in figure A.1. This result not only contradicts the result by Dolan et al. [7], but it also does not appear consistently, as the same settings give results for ALP-Strahlung that are in agreement with[7]. The effect was only observed in the $e^+e^-\gamma\gamma$ final state, and the rapid jump in the cross section, which does not trivially coincide with any mass that might be relevant in the process, is considered unphysical.
Further conversations with theory colleagues consolidated this assumption and gave helpful hints on how the effect might arise.[1] While the exact reason for this behaviour is not understood and might involve deeper studies of the ISR calculation code, fixing the renormalization and factorization energy scales to any value resolves the discontinuity and recovers results in agreement with Dolan et al.[7] Additional checks show that the difference of the exact numerical value for these scales does not influence the results when changing the scale energy between the $\Upsilon(4s)$ mass and the Z mass. Hence, the analysis uses the beam energy, resulting in stable cross-section curves.

[1] Thanks to Prof. Felix Kahlhöfer for the fruitful discussions on this topic, hinting us towards a solution.

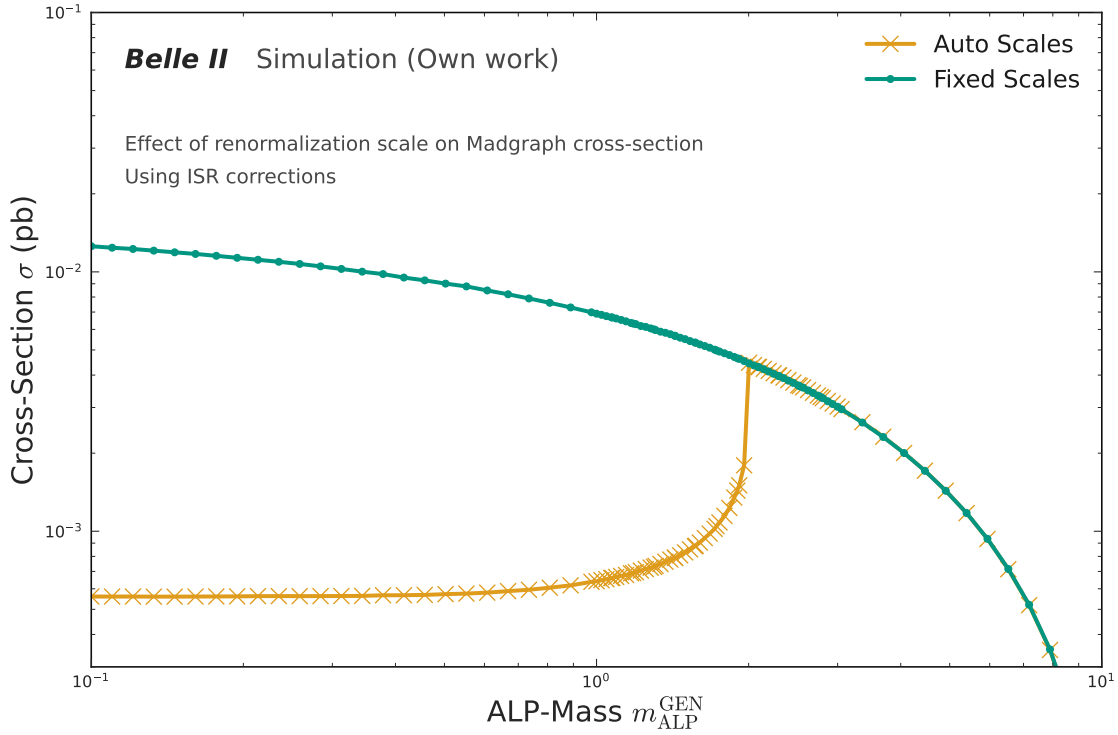


Figure A.1.: This figure shows the cross-section calculated by Madgraph over the ALP mass. Both lines are done with the corrections for the initial state radiation active. But the orange one is using the automated calculation of the renormalization and factorization scales, while the green one uses scales fixed to the beam energy.
The orange line exhibits a clear discontinuity around $2\,\mathrm{GeV}/c^2$, contradicting the results of Dolan et al. [7] and appearing unphysical. To mitigate this, the analysis uses the settings as the green line, resolving the discontinuity, matching the results of Dolan et al. [7]

A.1.2. Event Count Limit

The second inconvenience was identified when scaling up the analysis for classifier training. The used `MadGraph5_aMC@NLO` generator (Version 3.4.0) MC generator seems to have a limit on the maximal allowed number of events it generates of around 10^6 per execution. However, it builds an executable that can be used to repeat the generation and give a second set of additional MC events with identical settings.
To mitigate the limit, this analysis set up a tool to surpass 10^6 events by batching the requested number of events $n_{\text{Req.}}$ into

$$l = \left\lceil \frac{n_{\text{Req.}}}{10^6} \right\rceil, \tag{A.1}$$

batches with n_{Batch} events per batch calculated as

$$n_{\text{Batch}} = \left\lfloor \frac{n_{\text{Req.}}}{l} \right\rfloor . \tag{A.2}$$

With the new number of events per pass n_{Batch}, `MadGraph5_aMC@NLO` is called l times. Each of these calls results in an individual output file. As these files save the events in the Les Hoches Events (LHE) format, which is a version of eXtensible Markup Language (XML), the files can conveniently be merged together using highly optimized bash tools for text file handling. This allows for combining files with sizes in the order of 1 GB within seconds. Additionally, a check is performed that the output file is still properly formatted, and contains the proper number of events $n_{\text{Gen.}} = l \cdot n_{\text{Batch}}$. It is important to note that in general $n_{\text{Gen.}} \neq n_{\text{Req.}}$. However, the difference in the event count is on the order of the used batches l and is negligible[2]. Hence, all calculations, e.g,. signal efficiencies, will use the requested number of events instead of the actually generated one, and the term $n_{\text{Gen.}}$ is used for this outside of this discussion.

[2]For the used number of events in the analysis, the difference is 1 event for $n_{\text{Req.}} = 1.5 \cdot 10^6$ and 6 event for $n_{\text{Req.}} = 12.5 \cdot 10^6$

List of Acronyms

ALP Axion-Like Particle. 1–10, 25–29, 31, 33, 34, 37, 39, 43, 44, 57, 73, 74, 76, 79–81

ARICH Aerogel Ring-Imaging Cherenkov detector. 16, 17

AUC Area Under Curve. 53

basf2 Belle II Analysis Software Framework. i, 1, 11, 21, 22, 27–31, 33, 35, 36, 40

BDT Boosted Decision Tree. 40, 55

BES III Beijing Spectrometer III. 10, 74

BSM Beyond the Standard Model. 1, 3, 4, 12, 43

CDC Central Drift Chamber. 14, 16, 17, 20, 21, 29, 30, 33, 35, 44, 65

CMS Compact Muon Solenoid. 13

CP Charge-Parity. 4, 11, 13

CPU Central Processing Unit. 20

CPV Charge-Parity Violation. 4

DM Dark Matter. 3, 4

ECL Electromagnetic Calorimeter. 17, 20, 29, 33–35, 44, 65

FoM Figure of Merit. 52, 53

FPGA Field Programmable Gate Array. 20

HLT High Level Trigger. 20, 21

ISR Initial State Radiation. 27, 28, 81

KLM K_L and Muon Detector. 19, 20

L1 Level One. 20, 29

LHC Large Hadron Collider. 10, 27

LHE Les Hoches Events. 74, 83

MC Monte Carlo. 1, 21–23, 27, 29–31, 36, 39, 40, 45, 55, 61, 62, 65, 67, 80, 81, 83

MDST Mini Data Summary Table. 23, 30, 33

MVA MultiVariate Analysis. 35

NN Neural Network. 2, 43, 44, 51, 54, 79

PDF Probability Density Function. 66

PID Particle Identification. 13, 16, 17, 21, 22, 34, 35, 40, 41, 44

PXD Pixel Detector. 13, 14, 20, 21

QCD Quantum Chromodynamics. 4

QM Quantum Mechanics. 1

ROC Receiver Operating Characteristic. 53, 54

RPC Resistive Plate Chamber. 19

SM Standard Model. 1–5, 23, 25, 26, 43, 61, 66, 79

SVD Silicon Vertex Detector. 13, 21

TOP Time of Propagation Counter. 16, 17

VXD Vertex Detector. 13, 14

WandB Weights and Biases. 48

XML eXtensible Markup Language. 83

List of Figures

List of Tables

Bibliography

[1] J. Alwall et al. “MadGraph 5: going beyond”. In: *Journal of High Energy Physics* 2011.6 (June 2011). ISSN: 1029-8479. DOI: 10.1007/jhep06(2011)128. URL: http://dx.doi.org/10.1007/JHEP06(2011)128.

[2] L. Torvalds and J. Hamano. *Git*. Version 2.31. Distributed version control system. 2021. URL: https://git-scm.com/.

[3] Heidelbach A. and Schmitt F. *Alpstrahlung*. Version v1.2.3. Private repository. 2025. URL: https://gitlab.etp.kit.edu/aheidelbach/alpstrahlung.

[4] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021. URL: https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[5] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020. URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[6] P. W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[7] M. J. Dolan et al. “Revised constraints and Belle II sensitivity for visible and invisible axion-like particles”. In: (Jan. 2018). DOI: 10.1007/JHEP12(2017)094.

[8] F. Acanfora et al. “Fusing photons into diphoton resonances at Belle II and beyond”. In: (June 2024).

[9] V. C. Rubin, W. K. Ford Jr., and N. Thonnard. “Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc).” In: *Astrophysical Journal* 238 (June 1980), pp. 471–487. DOI: 10.1086/158003.

[10] C. Abel et al. “Measurement of the Permanent Electric Dipole Moment of the Neutron”. In: *Phys. Rev. Lett.* 124 (8 Feb. 2020), p. 081803. DOI: 10.1103/PhysRevLett.124.081803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.124.081803.

[11] R.D. Peccei and H. Quinn. “CP Conservation in the Presence of Pseudoparticles”. In: *Physical Review Letters - PHYS REV LETT* 38 (June 1977), pp. 1440–1443. DOI: 10.1103/PhysRevLett.38.1440.

[12] K. Choi et al. "Minimal flavor violation with axion-like particles". In: *Journal of High Energy Physics* 2017.11 (Nov. 2017). ISSN: 1029-8479. DOI: 10.1007/jhep11(2017)070. URL: http://dx.doi.org/10.1007/JHEP11(2017)070.

[13] F. Abudinén et al. "Search for Axionlike Particles Produced in e+e- Collisions at Belle II". In: *Physical Review Letters* 125.16 (Oct. 2020), p. 161806. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.125.161806.

[14] G. De Pietro, T. Ferber, and A. Heidelbach. *Search for ALPs in $e^+e^- \to \gamma\, a,\ a \to \gamma\gamma$ at Belle II.* Tech. rep. BELLE2-NOTE-PH-2025-048. Belle II internal note. Belle II Collaboration, 2025.

[15] Peicheng Jiang. "ALPs searches at BESIII". In: (May 2023).

[16] Belle II Collaboration. *Luminosity.* https://www.belle2.org/research/luminosity/. Accessed: 2025-09-17. 2025.

[17] H. J. Bhabha. "The scattering of positrons by electrons with exchange on Dirac's theory of the positron". In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 154.881 (1936), pp. 195–206. DOI: 10.1098/rspa.1936.0046. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1936.0046. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1936.0046.

[18] Belle II Collaboration. *Rendering of the Belle II Detector. Belle II website.* https://www.belle2.org/archives/. Accessed: 2025-09-15. 2025.

[19] T. Abe et al. "Belle II Technical Design Report". In: (Nov. 2010).

[20] B. Spruck et al. "Belle II Pixel Detector Commissioning and Operational Experience". In: *PoS* Vertex2019 (2020), p. 015. DOI: 10.22323/1.373.0015.

[21] R.L. Gluckstern. "Uncertainties in track momentum and direction, due to multiple scattering and measurement errors". In: *Nuclear Instruments and Methods* 24 (1963), pp. 381–389. ISSN: 0029-554X. DOI: https://doi.org/10.1016/0029-554X(63)90347-1. URL: https://www.sciencedirect.com/science/article/pii/0029554X63903471.

[22] V. Bertacchi et al. "Track finding at Belle II". In: *Computer Physics Communications* 259 (2021), p. 107610. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2020.107610. URL: https://www.sciencedirect.com/science/article/pii/S0010465520302861.

[23] H. Ikeda. *Development of the CsI(T') Calorimeter for the Measurement of CP Violation at KEK B-Factory.* Tech. rep. Nara Women's University, 1999.

[24] M. Ritter et al. "Belle II Conditions Database". In: *Journal of Physics: Conference Series* 1085.3 (Sept. 2018), p. 032032. DOI: 10.1088/1742-6596/1085/3/032032. URL: https://dx.doi.org/10.1088/1742-6596/1085/3/032032.

[25] E. Kou et al. "The Belle II Physics Book". In: (Sept. 2019). DOI: 10.1093/ptep/ptz106.

[26] Belle II Collaboration et al. *Search for a dark Higgs boson produced in association with inelastic dark matter at the Belle II experiment.* 2025. arXiv: 2505.09705 [hep-ex]. URL: https://arxiv.org/abs/2505.09705.

[27] S. Knapen et al. “Searching for Axionlike Particles with Ultraperipheral Heavy-Ion Collisions”. In: *Phys. Rev. Lett.* 118 (17 Apr. 2017), p. 171801. DOI: 10.1103/PhysRevLett.118.171801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.171801.

[28] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (July 2003), pp. 250–303. ISSN: 01689002. DOI: 10.1016/S0168-9002(03)01368-8.

[29] Sato, Y. et al. “Monte Carlo matching in the Belle II software”. In: *EPJ Web Conf.* 251 (2021), p. 03021. DOI: 10.1051/epjconf/202125103021. URL: https://doi.org/10.1051/epjconf/202125103021.

[30] P. Canal et al. *root-project/root: v6-36-04*. Version v6-36-04. Aug. 2025. DOI: 10.5281/zenodo.16944732. URL: https://doi.org/10.5281/zenodo.16944732.

[31] J.-F. Krohn et al. “Global decay chain vertex fitting at Belle II”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 976 (2020), p. 164269. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2020.164269. URL: https://www.sciencedirect.com/science/article/pii/S0168900220306653.

[32] Ferber T. *OrcaKinfit: A kinematic fitting library for Belle II*. https://github.com/tferber/OrcaKinfit. GitHub repository, accessed 2025-09-13. 2017.

[33] Apache Parquet Project. *Parquet File Format Specification*. Accessed: 2025-09-14. 2025. URL: https://parquet.apache.org/docs/file-format/.

[34] The pandas development team. *pandas-dev/pandas: Pandas*. Version v2.3.2. Aug. 2025. DOI: 10.5281/zenodo.16918803. URL: https://doi.org/10.5281/zenodo.16918803.

[35] Ritchie V. et al. *pola-rs/polars: Python Polars 1.33.1*. Version py-1.33.1. Sept. 2025. DOI: 10.5281/zenodo.17084198. URL: https://doi.org/10.5281/zenodo.17084198.

[36] M. Milesi, J. Tan, and P. Urquijo. “Lepton identification in Belle II using observables from the electromagnetic calorimeter and precision trackers”. In: *EPJ Web Conf.* 245 (2020), p. 06023. DOI: 10.1051/epjconf/202024506023. URL: https://doi.org/10.1051/epjconf/202024506023.

[37] F. Schmitt. *Improving Event Selection with Machine Learning Methods for the B to K a, a to gg Search at Belle II*. 2022.

[38] L. Biewald. *Experiment Tracking with Weights and Biases*. Software available from wandb.com. 2020. URL: https://www.wandb.com/.

[39] G. Punzi. “Sensitivity of searches for new signals and its optimization”. In: 2003.

[40] J. E. Gaiser. “Charmonium Spectroscopy From Radiative Decays of the J/ψ and ψ'”. PhD thesis. Stanford University, Aug. 1982.

[41] J. Eschle et al. “zfit: Scalable pythonic fitting”. In: *SoftwareX* 11 (Jan. 2020), p. 100508. ISSN: 2352-7110. DOI: 10.1016/j.softx.2020.100508. URL: http://dx.doi.org/10.1016/j.softx.2020.100508.

[42] J. Gäßler et al. “kafe2 – a Modern Tool for Model Fitting in Physics Lab Courses”. In: (2022). DOI: 10.48550/arXiv.2210.12768. URL: https://doi.org/10.48550/arXiv.2210.12768.

[43] G. Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0.

[44] M. Marinangeli et al. *scikit-hep/hepstats: v0.9.2*. Version v0.9.2. Mar. 2025. DOI: 10.5281/zenodo.14973858. URL: https://doi.org/10.5281/zenodo.14973858.