Wafer-Scale Stitched CMOS Pixel Sensors: Characterisation and Detector Performance Studies for ALICE ITS3

Gregor Hieronymus Eberwein

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Abstract

The ALICE experiment at the Large Hadron Collider at CERN will upgrade the innermost three layers of the currently installed Inner Tracking System (ITS2) to the ITS3 during Long Shutdown 3 (LS3), starting in 2026. The ITS3 introduces a groundbreaking detector design based on wafer-scale curved monolithic active pixel sensors. Each of the three silicon layers, thinner than 50 µm, will be bent around the beam pipe and supported by lightweight carbon-foam structures. The detector will be air-cooled. This configuration achieves an unprecedented material budget of just $X/X_0 \simeq 0.09\%$ per layer. This material reduction, combined with a decrease in the distance of the innermost tracking layer from the interaction point, results in a factor-of-two improvement in pointing resolution at transverse momenta below 10 GeV/c compared to the current detector. The ALICE core physics programme, requiring excellent secondary vertex reconstruction and low-mass tracking, will strongly benefit from this upgrade. To enable wafer-scale sensors up to 9.8×27 cm^2, the design employs stitching – a technique evaluated using the Monolithic Stitched Sensor (MOSS), a 14×259 mm^2 prototype fabricated in 65 nm CMOS technology. This work focuses on characterising MOSS in terms of fabrication yield, mechanical handling, and sensor performance. Custom hardware and software were developed for detailed characterisation, and a fault-diagnosis method was introduced to identify and resolve unexpected sensor failures. Finally, the performance of the final ITS3 in terms of pointing resolution and tracking efficiency was simulated assuming malfunctioning substructures on the sensor planes. Both qualitative and machine-learning-based approaches to optimising the geometrical arrangement of the layers are discussed, and the impact of reduced tracking efficiency on the Λ_c^+ benchmark observable is investigated.

Preface

Large-scale particle physics experiments at the LHC are inherently collaborative efforts, as constantly evolving discovery and measurement tools are built on decades of experience, teamwork, and fruitful discussions. My contributions during my doctoral research are outlined below in roughly chronological order.

For the ATLAS experiment, earning me authorship, I worked on the ITk upgrade. I developed an FPGA-based interlock system for the Oxford low-temperature pixel sensor module qualification setup, including design and construction of a relay PCB, temperature controller, graphical user interface, and a database interface for environmental logging. I wrote the ATLAS-internal document for interlock requirements for testing site-qualification [1]. I studied bump-bond delamination induced by thermal stress, revealing clear vendor differences that allowed for vendor processing improvements. I contributed to the resulting ATLAS internal bump-stress report [2], which was part of the final design review. This work was performed at Oxford, and is not documented in this thesis.

Over the following two years at CERN, I was part of the ITS3 project within the ALICE experiment. I took part in two test beam campaigns for the Analogue Pixel Test Structure (APTS) chip, a 65 nm prototype and technology-validation chiplet with results published in [3, 4]; this work is not further described here. My main contributions, described in this thesis, revolve around the characterisation and performance evaluation of the first-ever Monolithic Stitched Sensor (MOSS) prototype for high-energy physics. I was one of the first contributors to establishing the characterisation environment, taking on key responsibilities in its development:

- Contributed significantly to the custom tooling and procedures for handling, mounting, glueing, and bonding the MOSS sensor. I designed mechanical parts, including jigs, tooling components, chip covers and storage boxes, as well as 3D-printed parts. I developed and built a vacuum pump and control system, as well as a control system for semi-automated, motorised wafer positioning with a joystick interface – both based on microcontrollers with custom firmware. I proposed the use of a glue robot and implemented a high-precision glueing procedure, which I validated experimentally. I designed and built a custom UV-curing box to accelerate the curing process.

- Wrote the first Data Acquisition (DAQ) software interface for the Proximity boards, and contributed to the associated FPGA firmware. This enabled testing of the Proximity board functionality and on-board components such as DACs, LDOs, ADCs via I^2C and SPI communication protocols, and verification of pinouts. I identified bugs and solutions that were subsequently implemented by the electronics engineer, and a new board was produced.

- Served as one of the main contributors to the complete DAQ software framework for the MOSS functional characterisation, developing a versatile API that served as the foundation for subsequent testing scripts [5].

- Built a multiplexer-based impedance setup and developed an automated power ramping setup, incorporating a thermal camera. I also wrote corresponding data acquisition software for testing [5].

- Performed extensive measurements on wire-bonded MOSS sensors, including in-depth data analysis and iterative refinement of test procedures. I contributed to initial wafer-probing measurements.

- Proposed the inclusion of NTC temperature probes on the MOSS carrier PCB. I designed and implemented an external ADC board for the NTC readout to plug onto the FPGA readout board. I included an additional level-shifter board for feeding external trigger signals to the FPGA board and added the first version of the required firmware extension.

- Developed a failure analysis method to identify on-chip metal stack short-circuit faults. This method integrated semi-automated image processing, impedance measurements, and power ramp-up data, all correlated with chip design files. This approach successfully pinpointed the failure mechanism and enabled corrective action.

- Proposed the use of Focused Ion Beam Scanning Electron Microscopy for failure analysis, and validated the hypothesis by identifying defects consistent with the failure mechanism.

- Simulated the impact of dead areas in the final ITS3 sensor layout on detector performance. I introduced a machine learning based approach for optimising layer geometry and evaluated its feasibility. I also assessed the effect of reduced tracking efficiency on the Λ_c^+ benchmark channel.

The first MOSS yield and functional measurements, which I performed, were published as a key component of the ITS3 Technical Design Report [6]. A comprehensive article on the MOSS chip characterisation [7], and a dedicated paper – where I am lead author – on the metal stack analysis method I developed [8] are being prepared. Both papers are being reviewed at the time of writing.

Contents

1 Introduction **1**

2 LHC, ALICE, and ITS3 **5**

2.1 The Large Hadron Collider 5

2.2 The ALICE experiment 8

2.2.1 ALICE physics goals 8

2.2.2 ALICE Detector system 12

2.3 Particle tracking 15

2.4 ITS3 upgrade 19

2.4.1 Physics performance simulation 23

2.4.2 Λ_c^+ measurement 26

3 Monolithic Active Pixel Sensors, MOSS, and ITS3 sensors **31**

3.1 Interaction of particles with matter 31

3.2 Silicon as detector material 33

3.2.1 p-n junction and charge carrier transport 34

3.3 Monolithic Active Pixel Sensors – MAPS 37

3.4 Fabrication 39

3.4.1 Lithography 39

3.4.2 Chip interconnect technology and metal stack 40

3.4.3 Stitching 43

3.5 Monolithic Stitched Sensor MOSS 44

3.5.1 MOSS design and architecture 44

3.5.2 Metal stack and power grid 50

3.6 ITS3 sensor layout 53

4 MOSS Handling and Mounting **55**

4.1 Pick-up system 55

4.2 Mounting 60

4.2.1 MOSS carrier PCB 60

4.2.2 Chip mounting 62

4.3 Interconnection 67

5 MOSS Test Systems and Measurements **71**

5.1 Impedance measurement 71

5.1.1 Wafer probing 74

5.2 Power ramping with thermal camera analysis 75

5.2.1 Hotspot localisation with a thermal camera 77

5.2.2 Chip design correlation 81

5.3 Test system for functional tests 82

6 MOSS Characterisation and Data Analysis **93**

6.1 Measurement sequence 93

6.2 Impedance measurement results 95

6.2.1 Empiric definition of shorts at 30 Ω 95

6.2.2 Shorts on each wafer 96

6.2.3 Location of shorts on wafer level 101

6.2.4 Power net pairs affected by shorts 102

6.2.5 Observed vs. real number of shorts 103

6.3 Powering results 109

6.3.1 Classification of Half Units and powering yield 109

6.3.2 Power-on endpoint currents 115

6.3.3 Burn-through current and voltage 118

6.4 Post power impedance measurement results 119

6.5 Thermal camera results and short fault mechanism 123

6.6 Yield extrapolation 123

7 Failure and Root Cause Analysis **129**

7.1 Thermal camera measurement results 129

7.1.1 Short locations 130

7.1.2 Simultaneous shorts 131

7.1.3 Fluctuating shorts 133

7.2 Hypothesis formation on failure mode 134

7.3 [redacted] 136

7.3.1 [redacted] 139

7.4 FIB-SEM system 140

7.5 Sample analysis 142

7.6 Root cause, mitigation, and reliability 145

8 Effect of Dead Areas on the ITS3 Physics Performance and Optimisation Strategy 147
8.1 Simulation framework and procedure 147
8.1.1 Deadmap generation 148
8.2 ITS3 pointing resolution and tracking efficiency 150
8.2.1 Pointing resolution 150
8.2.2 Reconstruction efficiency 151
8.3 Effect of dead tiles on pointing resolution and tracking efficiency 151
8.3.1 Pointing resolution degradation 152
8.3.2 Reconstruction efficiency loss 155
8.3.3 Deadmap ranking 156
8.3.4 Qualitative two-condition approach for detector optimisation . 160
8.4 Optimisation of ITS3 layer geometry using an artificial neural network 162
8.4.1 Number of permutations of ITS3 layer arrangements 162
8.4.2 Neural network model training 163
8.4.3 Deadmap ranking with the artificial deep neural network 169
8.5 Effect of dead tiles on the Λ_c^+ reconstruction efficiency 175
8.5.1 Daughter particle tracking loss 176
8.5.2 Event generation and kinematics 176
8.5.3 Λ_c^+ efficiency loss 178

9 Conclusions and Outlook 183

Appendices

A Supplementary Figures 189

B Extended Yield Model with Clustering Factor 197

Abbreviations 201

References 203

1 Introduction

In the continuous pursuit of expanding the boundaries of our current understanding of particle physics, appropriate tools are essential for testing the limits of our theoretical models, challenging our state of knowledge, and uncovering new directions. Following the impressive success of the Standard Model (SM) as the governing theory of fundamental particle physics [9–11] – for decades serving as a guide for measurements and completed with the discovery of the Higgs boson in 2012 [12, 13] – the current data-driven era relies on instrumentation and tools providing the best information to advance the field. In addition to the many open questions not described by the SM – including neutrino masses, the origin of the matter–antimatter asymmetry, dark matter, and dark energy, to name a few – and continued testing of the SM's predictive power, the Quantum Chromodynamics (QCD) sector describing the strong interaction relies on an extensive measurement programme [14, 15]. The ALICE experiment at the Large Hadron Collider (LHC) at CERN is focused on the study of strongly interacting matter at extreme energy densities [16, 17]. Ultra-relativistic heavy-ion collisions at the LHC give access to the study of the Quark-Gluon Plasma (QGP) – a state of matter arising from asymptotic freedom (the running of the strong coupling $\alpha_s(Q^2)$), in which quarks and gluons, the carriers of the strong interaction, are deconfined. This is in contrast with ordinary nuclear matter, where quarks and gluons exist only in bound states such as the proton or neutron. To probe parameters

and characteristics of the QGP – which cannot be directly observed and must be inferred – ALICE requires precision measurements and reconstruction of particles down to very low momenta $p_T < 1\,\mathrm{GeV}/c$. To facilitate these measurements at the challenging high track multiplicities of heavy-ion collisions, novel silicon tracking and vertexing detectors are being employed and developed. Quoting Ian Shipsey, *instrumentation is the great enabler* of science, with advanced silicon detectors for high-energy physics experiments being a crucial tool among many [18].

The inner tracking system of the ALICE experiment was upgraded to the Inner Tracking System 2 (ITS2) in 2021 [19]. With 10 m^2 sensitive area, it is the largest silicon pixel detector built for a high-energy physics experiment, based on Monolithic Active Pixel Sensors (MAPS). With the reduction of the material budget (to $X/X_0 = 0.36\%$ per layer in the inner barrel), decrease in pixel size, and reduction of the distance between the innermost layer and the interaction point, a factor of 3–5 improvement in pointing resolution was achieved [20].

The proposed Inner Tracking System 3 (ITS3) will improve on this design, replacing the innermost three layers of the ITS2 [6, 21]. This novel vertex detector is based on wafer-scale stitched monolithic active pixel sensors and is fabricated using a commercial 65 nm CMOS imaging process. The wafer-scale sensors, of up to 27×9.8 cm^2 in size, are thinned to a thickness below 50 μm, allowing them to be bent into a (half-)cylindrical shape, and placed even closer around the beam pipe. The resulting cylindrical geometry is sufficiently rigid that only an ultralight carbon-foam support is needed when using air cooling, reaching $X/X_0 \simeq 0.09\%$ per layer, and achieving another factor-of-two improvement in pointing resolution for transverse momenta $p_T \lesssim 10\,\mathrm{GeV}/c$.

An extensive R&D programme is dedicated to the development of three key areas pioneered in the ITS3 project: (a) use of 65 nm CMOS technology for MAPS, (b) bending of silicon, and (c) stitching. Stitching is a manufacturing technique that allows the fabrication of silicon sensors larger than the design reticle size. It is realised by designing structures with periodic boundaries, and electrically interconnecting

the repeated structures on-chip across the stitching boundaries at the metal-stack level. A prototype sensor, the Monolithic Stitched Sensor MOSS, was designed to evaluate the feasibility of this technique for high-energy physics, with the primary goals of determining the crucial yield parameter and understanding the factors and performance characteristics which contribute to a successful design [22]. The MOSS sensor characterisation – including the development of novel tooling, procedures, and measurement methods – is the focus of this work, complemented by simulation of the final detector performance accounting for imperfect sensor layers. Novel technologies require extensive testing to identify limitations and the root causes of associated failures, enabling their mitigation in later iterations toward the final design.

The development of wafer-scale, flexible MAPS for ultra-low material budget tracking systems is well advanced, offering unique potential for future detector systems. Beyond its first application in the ALICE ITS3 vertex detector, the technology will be adopted for the silicon tracking system of the ePIC detector at the Electron–Ion Collider (EIC) [23]. The proposed ALICE 3 detector's vertex system will rely on flexible, radiation-hardened, and low-material-budget MAPS [15]. Future e^+e^- collider experiments, such as the proposed Future Circular Collider (FCC-ee), will require low-mass precision trackers to fully exploit the physics potential, sharing similar vertex-detector requirements [24]. As the first of its kind, ITS3 will pave the way for the development of exciting new tracking systems capable of unprecedented performance, enabling previously out-of-reach measurements.

This work is structured as follows. Chapter 2 introduces the LHC, discusses the ALICE physics programme and detector concept, before describing the principles for high spatial resolution tracking. The last section discusses the ITS3 upgrade and the expected improvement using the Λ_c^+ benchmark measurement. In Chapter 3, the working principle of silicon as a particle detector is discussed, monolithic active pixel sensors – including fabrication techniques relevant to this work – are introduced, the MOSS sensor is described in detail, and the conceptual ITS3 sensor layout is

referenced. Chapter 4 describes the development and use of custom tooling and procedures required for the successful handling, mounting, and interconnection of the MOSS sensor. In Chapter 5, the characterisation setups that were developed in this work and the corresponding measurements performed are discussed. The results of the measurements are discussed in Chapter 6, focusing on the observed short-circuit failure and its origin, and concluding with a discussion of the yield implications for future devices. A detailed discussion of the root cause and fault mechanism of the observed failure is given in Chapter 7[1]. In Chapter 8, studies on the ITS3 physics performance are described, assessing the impact of malfunctioning sensor sub-parts in the final system, discussing practical implications for the detector construction, and investigating the feasibility of a machine-learning-based geometry optimisation. The effect of reduced tracking efficiency on the Λ_c^+ benchmark observable is estimated. Finally, conclusions and outlook are given in Chapter 9.

In this spirit:

Measure what is measurable and make measurable what is not so.
– Galileo Galilei

[1]Given the nature of its contents, Chapter 7 might not be available to you at the time of reading. Contact the author or Bodleian Libraries, Oxford, for further information.

2
LHC, ALICE, and ITS3

This chapter introduces the Large Hadron Collider (LHC), A Large Ion Collider Experiment (ALICE) and its physics goals and detector systems, before discussing general considerations on tracking detectors. Finally, the ALICE Inner Tracking System 3 (ITS3) upgrade and the Λ_c^+ benchmark measurement are discussed.

2.1 The Large Hadron Collider

With a circumference of 26.7 km, the Large Hadron Collider (LHC) at the European Organisation for Nuclear Research (CERN) is the largest, most powerful hadron collider ever built [26]. Located between 45 m and 170 m below the surface, and crossing the border between France and Switzerland close to Lake Geneva, two counter-propagating hadron beams are held on-path by a superconducting magnet lattice and brought to collision at four Interaction Points (IP). The accelerator complex at CERN is depicted in Figure 2.1, with the LHC the last piece in the accelerator chain. The first collisions were recorded at the LHC in 2009, and the systems are continuously upgraded between data-taking periods, known as 'Runs'. The current Run 3 started in 2022 and will last until 2026, when the next upgrade period, the three-year Long Shutdown 3 (LS3), is planned. Parameters for Run 3 are quoted unless otherwise noted. In the current configuration, for proton-proton collisions (pp),

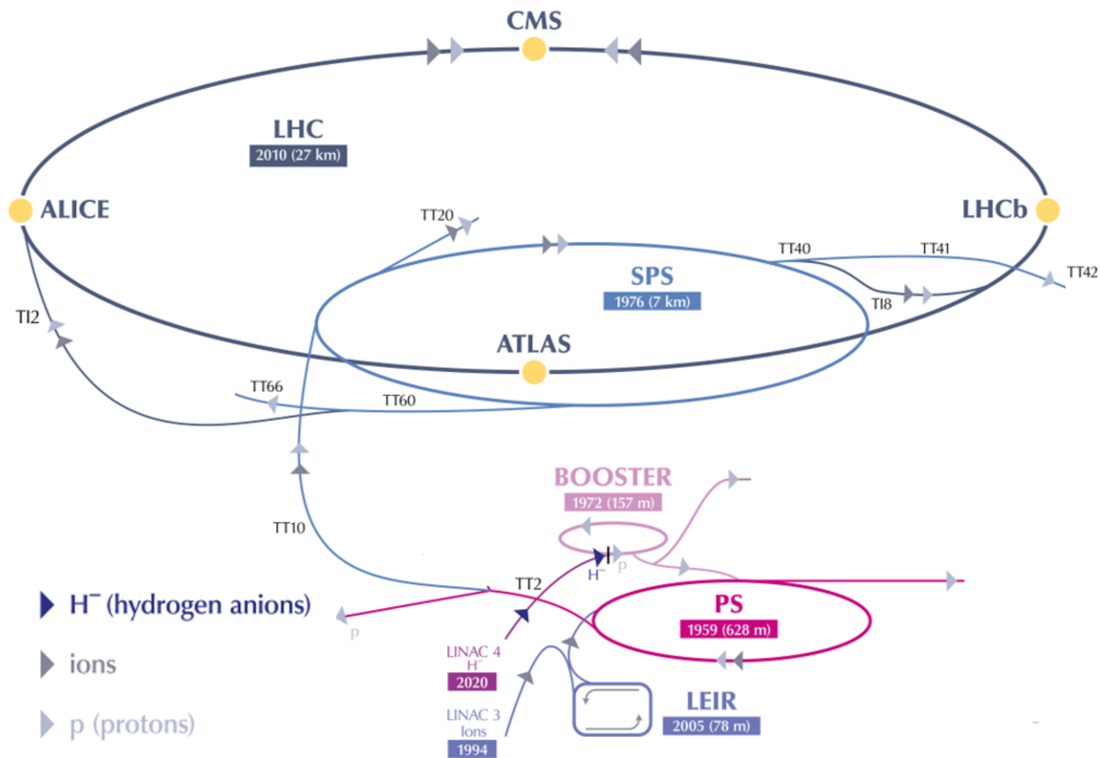


Figure 2.1: The LHC accelerator complex at CERN. Adapted from [25]. The four LHC interaction points are indicated as yellow dots, where the four large experiments are located (see text). Unconnected or open transfer lines (TT) indicate paths to additional experiments and are omitted for clarity.

the LHC injection chain starts with hydrogen anions (H^-) from an ion source that are accelerated with the Linear Accelerator (LINAC 4) up to 160 MeV energy [27]. The H^- anions are stripped of their two electrons at injection to the Proton Synchrotron Booster (PSB), where they are accelerated to 1.4 GeV energy, before being accelerated to 26 GeV by the Proton Synchrotron (PS). At this stage, protons are 'bunched' and spaced for further acceleration in the Super Proton Synchrotron (SPS) up to the LHC injection energy of 450 GeV. Two transfer lines connect to the LHC and inject the beams into the two circular beam lines. Each beam consists of 2,808 bunches (of which 2,464 are filled), with 25 ns spacing between bunches. Each bunch contains approximately 1.6×10^{11} protons. The two counter-propagating beams are accelerated by eight radio frequency cavities each. The beams are kept on their circular path by 1,232 superconducting dipole magnets, each 14.3 m long, operating at a nominal field of 8.33 T. In addition, more than 4,800 higher-order corrector magnets are used to

maintain and control the beam quality. Apart from protons, ions (e.g. Pb-nuclei) are accelerated by the LINAC 3 and Low Energy Ion Ring (LEIR) before being coupled into the PS, where they follow the common LHC injection path (albeit with an adapted filling scheme). Beams are brought to collision at four interaction points, achieving peak centre-of-mass energies of $\sqrt{s} = 13.6$ TeV and $\sqrt{s_{NN}} = 5.36$ TeV (per nucleon-pair) for pp and Pb-Pb collisions, respectively [28, 29]. At each interaction point, one of the four large experiments, ATLAS (A Toroidal LHC ApparatuS) [30], CMS (Compact Muon Solenoid) [31], ALICE (A Large Ion Collider Experiment) [16], and LHCb (Large Hadron Collider beauty experiment) [32] is located. ATLAS and CMS are considered multi-purpose detectors at the energy frontier, with the most prominent measurement the observation of the Higgs boson in 2012 [12, 13]. LHCb specialises in decay studies of b and c hadrons, investigating the matter-antimatter asymmetry with a unique forward spectrometer detector layout. ALICE is a purpose-built detector studying the properties of strongly interacting matter, in particular the Quark-Gluon Plasma (QGP), as further discussed below.

The number of expected events N_{exp} in time interval t is calculated as the product of the cross-section of interest σ_{exp} and the integrated luminosity $\mathscr{L}_{int}$:

$$N_{exp} = \sigma_{exp} \cdot \mathscr{L}_{int} = \sigma_{exp} \cdot \int_0^t \mathscr{L}(t')\, dt' \tag{2.1}$$

The instantaneous machine luminosity $\mathscr{L}$, commonly stated in units of cm^{-2} s^{-1}, is written as

$$\mathscr{L} = f_{coll} \frac{N_1 N_2}{4\pi\sigma_x^*\sigma_y^*} \mathscr{F} = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi\epsilon_n\beta^*} \mathscr{F} \tag{2.2}$$

where $N_1 N_2 = N_b^2$ is the number of particles per bunch, and the collision frequency is $f_{coll} = n_b f_{rev}$, with n_b bunches at a revolution frequency f_{rev}. The Root Mean Square (RMS) transverse beam sizes at the interaction point are characterised by $\sigma_x^*\sigma_y^* \simeq \epsilon_n\beta^*\gamma_r^{-1}$. Here, β^* is the beta function at the interaction point describing the final focus, γ_r is the relativistic Lorentz factor, and ϵ_n is the normalised emittance. The factor $\mathscr{F} \lesssim 1$ describes the reduction in instantaneous luminosity due to geometrical effects

such as the crossing angle, finite bunch length, and focusing of the two beams [26, 33]. In recent runs, peak instantaneous luminosities above $2.1 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ have been achieved for pp collisions, twice the LHC design value (recent values at [34]). The total integrated luminosity is commonly reported in inverse (femto)barns, e.g. $1\,\mathrm{fb}^{-1} = 10^{-39}\,\mathrm{cm}^{-2}$. In the ongoing Run 3 (up until July 2025), ATLAS recorded a total integrated luminosity of 206 fb^{-1} at $\sqrt{s} = 13.6$ TeV, with a mean number of interactions per bunch crossing (called pile-up) of $\langle\mu\rangle = 54$ [35] at a nominal 40 MHz interaction rate. In ALICE, during Pb-Pb collisions in Run 3 (up to July 2025), data were taken at a reduced ('levelled') instantaneous luminosity of $6.4 \cdot 10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and a nominal interaction rate of 50 kHz. With a maximum of 1,240 bunches per ring at 50 ns spacing (in the 2024 configuration), a total integrated luminosity of 3.08 nb^{-1} at $\sqrt{s_{NN}} = 5.36$ TeV was recorded [36, 37]. ALICE also records pp, p-Pb, and special-run nucleon collision data (such as Xe-Xe [38], or O-O [39]). Because of the reduced interaction rate, lower luminosity, and shorter overall run time compared to experiments such as ATLAS and CMS, the radiation hardness requirements of ALICE – particularly for detectors in the central region – are significantly less demanding. For example, the expected displacement damage in silicon sensors is lower by about three orders of magnitude. This more benign environment enables the development and deployment of novel, cutting-edge technologies, such as those planned for the future ITS3 detector discussed in this work.

2.2 The ALICE experiment

In this section, after a brief and qualitative introduction to quantum chromodynamics and ALICE physics goals and observables most relevant to this work, the ALICE detector – with a focus on the tracking system – will be described.

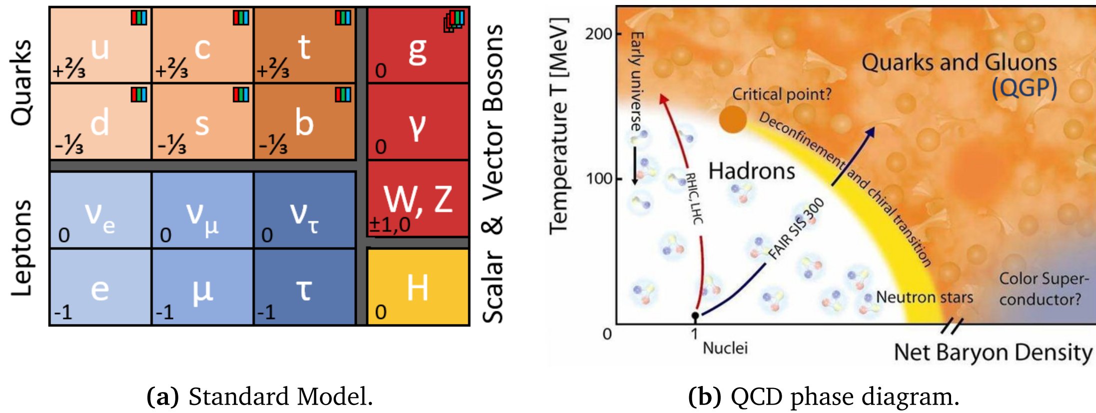
2.2.1 ALICE physics goals

ALICE is designed to study the strong interaction sector of the Standard Model of Particle Physics (SM), described by Quantum Chromodynamics (QCD) [11, 17, 33, 40, 41]. A simplified overview of the fundamental SM constituents is illustrated in

Figure 2.2a. Three generations of particles, split into leptons and quarks, are classified as fermions with spin $\frac{1}{2}$ (their charge-conjugated partners – antiparticles – shall be considered included). Vector bosons, or gauge bosons, with spin 1, represent the force carriers. The single spin-0 scalar boson – the Higgs boson – allows for the bare mass generation of the other massive fundamental particles via the Higgs mechanism. Fundamental forces, excluding gravity, are mediated by the SM vector bosons: the massless photon mediates the electromagnetic interaction, $W^{\pm}$, Z bosons mediate the weak interaction, and gluons mediate the QCD strong interaction.

Particles interact with the strong force if they carry colour charge, as indicated in Figure 2.2a. Gluons themselves carry colour charge and are, therefore, self-coupling. Strongly interacting hadrons, such as protons and neutrons, are composite particles made of quarks. They carry integer electric charges and have a net colour charge of zero. They are further classified as either mesons, which consist of a quark–antiquark pair and have integer spin, or baryons, which consist of three quarks and have half-integer spin. Colour-charge, as an additional quantum number, allows Pauli-excluded hadron states with three identical flavour-spin quarks, such as e.g. the $\Delta^{++}(1232) = uuu$.

A distinct feature of QCD is the behaviour of the running of the strong coupling $\alpha_s(Q^2)$, which describes the strength of the interaction as a function of the momentum transfer scale Q^2. Qualitatively, α_s is large at small Q or large distances, and decreases approximately as $1/\ln Q^2$ at large Q or small distances. This results in 'asymptotic freedom' of strongly interacting constituents at high energies and 'colour confinement' at large distances. As a consequence, at high densities or temperatures, quarks and gluons become quasi-free and form a deconfined phase of matter termed the Quark-Gluon-Plasma (QGP) [42]. The characteristic phase diagram of nuclear matter is shown in Figure 2.2b, where the net-baryon density n_B, is zero for an equal number of baryons and antibaryons. Lattice QCD predicts that, at zero net baryon density, the transition between hadronised matter and the QGP is a smooth crossover. However, the nature of the phase transition at higher net baryon densities remains an open question [44]. It is believed that the universe, between approximately 10 ps and 10 µs



(a) Standard Model. **(b)** QCD phase diagram.

Figure 2.2: (a) Standard Model constituents with electrical and colour charges indicated. (b) Phase diagram of nuclear matter. From [43].

after the Big Bang, existed in the form of a QGP. Similar conditions – characterised by low net baryon densities and extremely high temperatures – are recreated in heavy-ion collisions at the LHC [42].

ALICE is focused on the study of the properties of the QGP and its temporal evolution. During the pre-equilibrium phase, after the collision of the ions and particle formation from hard scattering processes, the system thermalises and the QGP forms. The following expansion of the QGP system is well described by hydrodynamics [45]. At decreasing density and temperature, hadronisation – the phase transition from a deconfined QGP to a hadronic state – occurs. First, the 'hadron gas' forms, where inelastic collisions persist until the 'chemical freeze-out', fixing particle abundances. Elastic scatterings then dominate until the 'kinetic freeze-out', fixing the momentum distributions, and the hadrons then stream freely from the interaction point Given the small size $O(\text{fm})$ due to confinement and short lifetime $\lesssim 10\,\text{fm}/c$ of the QGP, it can not be directly observed, but its existence and properties must be inferred via appropriate QGP probes [46].

A range of QGP signatures and corresponding probes exist, as described, e.g., in [47] – selected examples are provided here. Probes, which do not interact strongly with the QGP, such as direct photons or dileptons (e^+e^- or $\mu^+\mu^-$), serve as penetrating probes of the medium. Low invariant-mass di-electron pairs, for example, as produced

by internal conversion of virtual photons, are used for the measurement of electromagnetic radiation produced by the thermal QCD medium, providing access to its temperature and space-time evolution. A detector with acceptance for the lowest possible e^+e^- invariant masses and transverse momenta, down to at least $M_{ee} \sim p_{T,ee} \sim T \sim 150$ MeV – with electron detection down to $p_T < 100$ MeV/c – is required. However, suppression of thermal dilepton production and a large background from semi-leptonic charm decays and photon conversions before the first track measurement point make this measurement very challenging [47, 48]. These difficulties underline the requirement for very low material budget in the innermost part of the ALICE detector for reduction of photon conversions and improved low-p_T tracking capabilities (see also Section 2.3). The measurement will, therefore, benefit from the ITS3 upgrade (see Section 2.4), with improved low-p_T reconstruction efficiency of photon conversions to reduce the combinatorial background, and improved pointing resolution enabling the efficient tagging of electrons from semi-leptonic charm decays [21].

Heavy-flavour physics concerns the study of hadrons containing charm or beauty quarks and their use as probes of the collision system and QGP development [49]. Heavy-flavour quarks – due to their large masses – are primarily produced in hard-scattering processes before the formation of the QGP, and given their lifetime larger than the QGP itself, they experience the full evolution of the medium. Light-flavour hadrons, in contrast, are light enough that they can be created within the QGP, and their abundance is not constant over time. Charm and beauty quarks interact with the medium constituents, preserving, however, their flavour identity during energy exchanges. They are used as markers, resolving medium constituents and allow to build a connection between microscopic local partonic interactions and global medium characteristics – in particular, the medium's transport properties [50]. For example, the nuclear modification factor $R_{AA}(p_T)$, the scaled ratio of Pb-Pb over pp cross sections of an observable, is considered a sensitive observable for the study of the interaction of hard partons with the medium [49]. Furthermore, insights into hadronisation processes can be gained from production yields of heavy-flavour hadrons such as the Λ_c^+, as further discussed in Section 2.4.2. As heavy-flavour hadrons often have short

lifetimes $O(c\tau \sim 100\,\mu\text{m})$, excellent secondary vertex reconstruction capabilities are required for efficient background rejection.

ALICE is also primed for measurements of hypernuclei. Hypernuclei contain at least one strange baryon such as the Λ, substituting a proton or neutron. Hypertriton $^{3}_{\Lambda}\text{H}$ is the lightest known hypernucleus as the bound state of one proton, one neutron and one Λ. The measurement in ALICE is performed via the reconstruction of mesonic decay channels, e.g. $^{3}_{\Lambda}\text{H} \rightarrow ^{3}\text{He} + \pi^{+}$. Production yield and lifetime measurements offer unique insights into the formation and hyperon-nucleon interaction mechanisms. Hypernuclei decaying before the innermost detector layer of ALICE rely on efficient primary ^{3}He background rejection – improved with better impact parameter resolution (see Section 2.3). For example, the hyperhelium $^{4}_{\Lambda}\text{He} \rightarrow ^{3}\text{He} + \text{p} + \pi^{-}$ decay will strongly benefit from the inner tracker upgrade discussed in this work, with a signal-to-background ratio improvement by more than a factor of 3[1]. Additionally, measurements or new constraints of not-yet observed c-deuterons – nuclei containing charm quarks such as a potential neutron-Λ_c^{+} bound state – will improve [50].

Overall, a wide range of measurements relies on low-momentum track reconstruction, posing an interesting challenge to detector design that requires innovative technologies to advance performance further.

2.2.2 ALICE Detector system

The ALICE detector, shown in Figure 2.3, consists of a central barrel, which measures hadrons, electrons, and photons, and a forward muon spectrometer. The central part is embedded in a water-cooled, room-temperature solenoid magnet with a 0.5 T field strength, inherited from the L3 experiment at LEP, covering polar angles from 45° to 135°. In the forward spectrometer, the dipole magnet, with a horizontal field perpendicular to the beam axis, has a field integral of 3 Tm in the forward direction. The barrel part, from the inside out, consists of an Inner Tracking System (ITS), a cylindrical Time Projection Chamber (TPC), Particle Identification Detectors

[1] The first measurements of A=4 (anti)hypernuclei at the LHC with ALICE were recently published with first evidence for $^{4}_{\Lambda}\overline{\text{He}}$ at 3.5 σ significance and $^{4}_{\Lambda}\overline{\text{H}}$ at 4.5 σ significance [51].

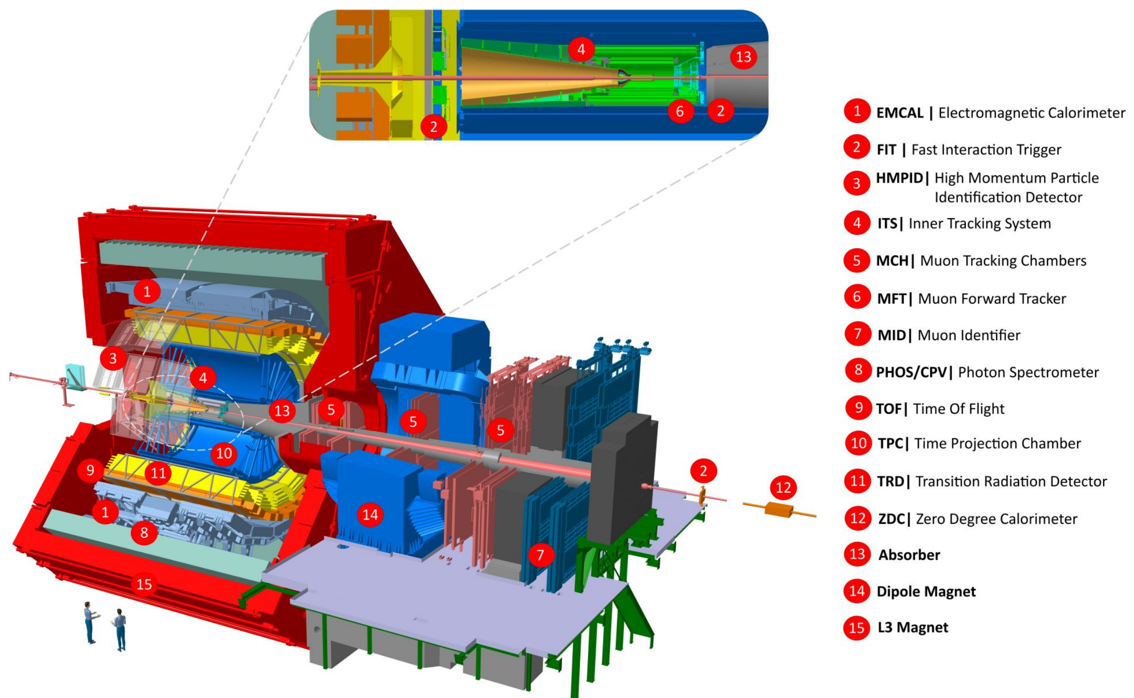


Figure 2.3: The ALICE detector in Run 2 configuration. Its overall dimensions are $16 \times 16 \times 26\ \mathrm{m}^3$ with a mass of approximately 10,000 t. After [52].

(PID) from Time-of-Flight (TOF), Transition Radiation Detector (TRD), and the High-Momentum Particle Identification Detector (HMPID) based on Ring Imaging Cherenkov (RICH) technology. These are followed by two electromagnetic calorimeters: the Electromagnetic Calorimeter (EMCAL) and the Photon Spectrometer (PHOS). In the forward muon region, covering polar angles from 2° to 9°, 14 layers of Muon Forward Tracking Chambers (MCH) and Muon Identification Detector (MID) chambers are installed following the Muon Forward Tracker (MFT) and an arrangement of absorbers. An additional set of smaller detectors (ZDC, PMD, FMD, T0, V0) is used for event characterisation and measurements at small angles from the beam axis [16]. Detector systems are continuously upgraded, and the configuration after LS2 and beyond operates in a continuous readout mode at a 50 kHz Pb-Pb interaction rate [52].

The upgraded TPC [53], which surrounds the central ITS (further discussed below), is a large-volume gas chamber that provides precise tracking and, importantly, also PID information. It consists of a cylindrical drift volume of 88 m^3 with a Ne-CO_2-N_2 gas mixture, a central high voltage electrode at 100 kV (400 $\mathrm{V\,cm^{-1}}$) and a maximal

drift length of 250 cm. Traversing charged particles leave a track of ionisation, and liberated e^- drift towards the end plates where Gas Electron Multiplier (GEM) detectors are used for readout.

Inner Tracking System

High-resolution vertex reconstruction and particle tracking close to the IP are performed with the ITS. The currently installed ITS2, as shown in Figure 2.4, is an upgrade installed in LS2 (2019–2022) [52, 54].

It is a silicon-only inner tracker based on Monolithic Active Pixel Sensors (MAPS, see also Chapter 3), with a total of 12.5 billion pixels and about 10 m^2 instrumented area. The ALICE Pixel Detector (ALPIDE) chip [55] with 512×1024 pixels, dimensions of 1.5×3.0 cm and a thickness of 50 μm was developed for use in ITS2. The Inner Barrel (IB) consists of 3 sensor layers (432 ALPIDE chips), with the Outer Barrel (OB) comprising 4 layers of sensors (23,688 ALPIDE chips). Sensors are mounted on an ultra-light carbon fibre support structure and water-cooled. It is the largest MAPS-based pixel detector built for a collider experiment [56].

Compared to the previous ITS, the material budget was significantly reduced from $X/X_0 = 1.14\%$ per layer to $X/X_0 = 0.36\%$ per layer (IB), while increasing the readout rate from 1 kHz to 50 kHz for Pb-Pb collisions. Together with a reduction of the innermost layer radius from 39 mm to 23 mm (reduction of beam pipe radius from 28 mm to 18 mm), the pointing resolution (see Section 2.3) is improved by a factor 3 in the transverse plane, and a factor 6 in the longitudinal plane.

These improvements are crucial for suppressing backgrounds in dielectron measurements and for reconstructing decays of heavy-flavour hadrons, enabling previously inaccessible measurements of heavy-flavour production. They are also essential for reconstructing very low-momentum tracks (e.g. $p_T \sim 80$ MeV for pions [57]) and secondary vertices.

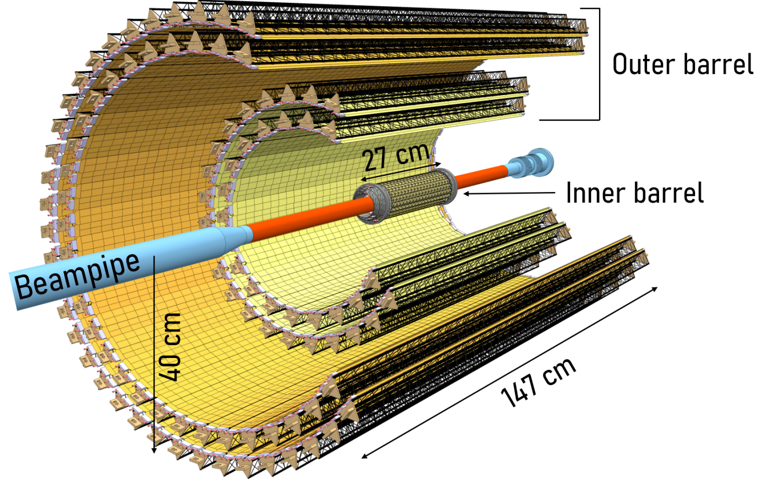


Figure 2.4: Installed upgrade of the Inner Tracking System (ITS2) [54].

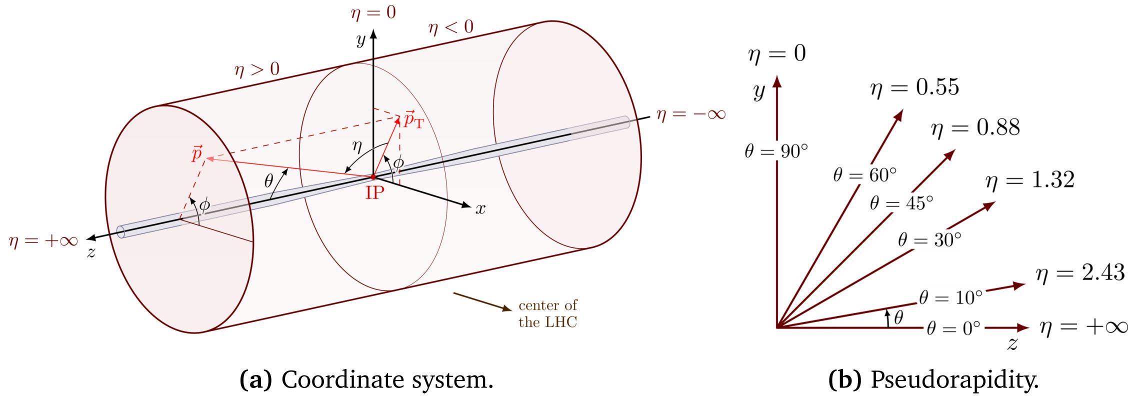
2.3 Particle tracking

Measures of particle tracking performance and variables for optimising the impact-parameter (pointing) resolution are introduced here. The coordinate system used is shown in Figure 2.5a. For particle tracks, cylindrical coordinates are preferentially used, with the beam line aligned along the z-axis. The particle's three-momentum $\mathbf{p}$ has a longitudinal component $p_L = p_z$ (along the beam) and transverse momentum magnitude $p_T = \sqrt{p_x^2 + p_y^2}$. The pseudorapidity, η, is defined as

$$\eta := -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{|\mathbf{p}| + \mathrm{p_L}}{|\mathbf{p}| - \mathrm{p_L}}\right) \tag{2.3}$$

where θ is the polar angle between $\mathbf{p}$ and the positive z-axis. In the high-energy limit $E \sim p$, the rapidity $y = \frac{1}{2}\ln\left(\frac{E+p_L}{E-p_L}\right)$ becomes the pseudorapidity [58]. Rapidity is particularly useful because differences in y are invariant under Lorentz boosts along the beam direction.

Charged particles of charge q in a magnetic field $\mathbf{B}$, such as the one generated by the L3 solenoid in ALICE, experience the Lorentz force $\mathbf{F} = q\,(\mathbf{v} \times \mathbf{B})$. From the measurement of parameters of the charged particle's helical trajectory in a magnetic field (here, a homogeneous field along the beam axis is assumed), its momentum can be calculated. For a reconstructed radius of curvature R, the transverse momentum p_T is given by [59]:



(a) Coordinate system. **(b)** Pseudorapidity.

Figure 2.5: (a) Coordinate system definition for experiments at the LHC. (b) Examples of pseudorapidity and corresponding angle from the beam axis. Adapted from [60].

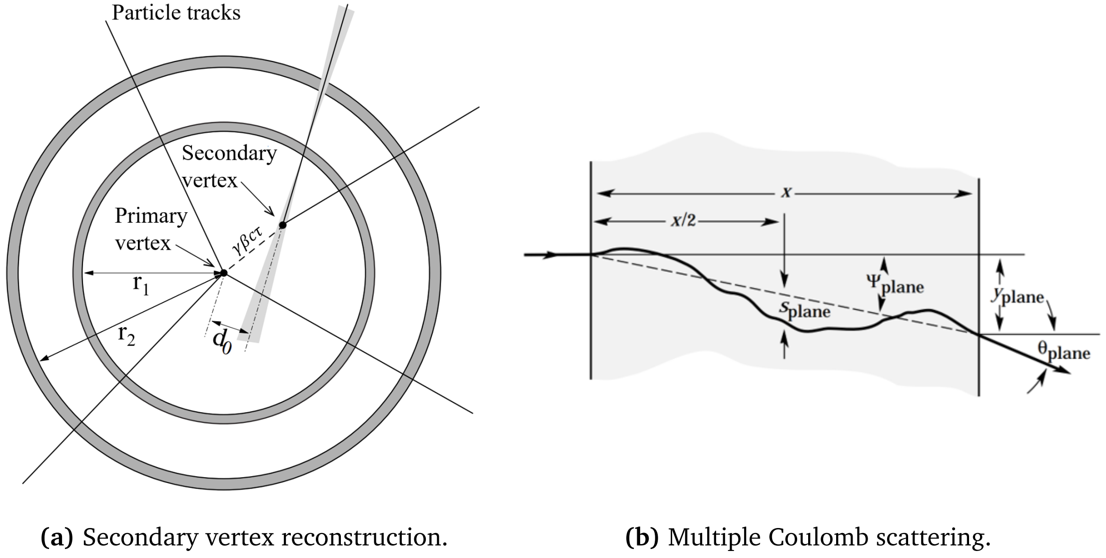
$$p_T = |q| B R \quad \rightarrow \quad p_T [\mathrm{GeV}/c] = 0.3\,|z|\,B\,[\mathrm{T}]\,R\,[\mathrm{m}] \tag{2.4}$$

with $z = q/e$ the particle charge in units of the elementary charge, and the velocity of light $c \approx 0.3 \cdot 10^9\,\mathrm{m/s}$.

An important measure of the performance of the inner tracking system is its ability to resolve secondary vertices. Short-lived particles, produced at the Primary Vertex (PV) in the collision and with lifetime τ, may decay far before reaching the first detection layer after a decay length $l = \gamma\beta c\tau$. The impact parameter d_0 is defined as the perpendicular Distance of Closest Approach (DCA) between the primary vertex and a given particle track, as illustrated in Figure 2.6a. If the impact parameter resolution (or pointing resolution) is sufficient, the secondary vertex is resolved by measuring ≥ 2 tracks, and the decay length l can be computed. The contributions to the impact parameter resolution σ_{DCA} can be discussed with a simplified model of two detector layers.

Charged particles traversing a medium scatter in the Coulomb fields of nuclei according to the Rutherford cross section. The Multiple Scattering (MS) deflection θ^{RMS}_{plane} (the RMS of the Gaussian describing the angular scatter distribution) by many small-angle scatterers, well described by the theory of Molière [61], is approximated by

$$\theta_{MS} = \theta^{RMS}_{plane} = \frac{13.6\,\mathrm{MeV}}{\beta c p}\, z \sqrt{\frac{X}{X_0}} \left(1 + 0.038 \ln \frac{X}{X_0}\right) \tag{2.5}$$



(a) Secondary vertex reconstruction. **(b)** Multiple Coulomb scattering.

Figure 2.6: (a) Illustration of particle tracks originating from the primary vertex and from a short-lived particle decaying at the secondary vertex. The transverse (xy-plane) view of a tracking detector with the first layer (and/or beam pipe) at r_1 and the second layer at r_2 is shown. The shaded area indicates the resolution degradation due to multiple scattering. Adapted from [59]. (b) Multiple Coulomb scattering of a particle traversing material of thickness x [33].

with p, βc, z the momentum, velocity, and charge number of the incident particle, respectively. The traversed medium thickness is written in radiation lengths X/X_0, with $X = X_0$ defined as the mean path length over which a high-energy electron loses all but $1/e$ of its energy by Bremsstrahlung [33]. The term X/X_0 is a common descriptor of material budget, and minimisation yields a reduction in MS-induced deflection of tracks. Especially for low momenta p, the contribution becomes highly significant ($\theta_{MS} \propto 1/p$)[2]. The MS contribution to the impact parameter resolution for the innermost layer radius r_1 becomes

$$\sigma_{DCA}^{MS} \sim \theta_{MS}\, r_1 \,. \tag{2.6}$$

The geometrical contribution to the impact parameter resolution is illustrated in Figure 2.7, which schematically shows the beam axis, beam pipe, and two detector layers at radii r_1 and r_2 with intrinsic spatial resolutions σ_1 and σ_2, respectively. The

[2] An interesting measurement of the ITS2 impact parameter resolution at $0.2 < p < 0.3$ GeV/c, sensitive to individual electronic components and mechanical structures on a single ALPIDE sensor-level in the innermost tracking layer, can be found in [62].

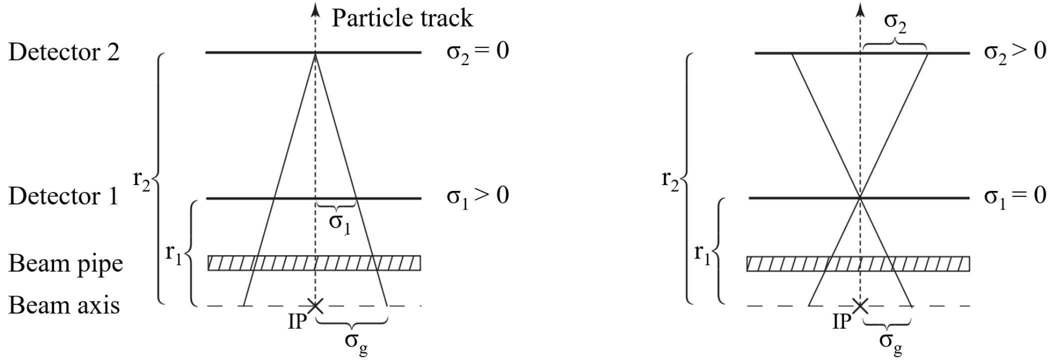


Figure 2.7: Effect of detector resolution on the measured geometrical position resolution σ_g. A perfect detector 2 with $\sigma_2 = 0$ is assumed on the left, with a perfect detector 1 with $\sigma_1 = 0$ on the right-hand side. Adapted from [59].

intrinsic spatial resolution σ_i of a pixel detector (such as ALPIDE, and sensors discussed in Chapter 3.5) with binary readout (hit/no-hit) and pixel pitch d_p is given by [63]:

$$\sigma_i^2 = \frac{1}{d_p} \int_{-\frac{d_p}{2}}^{+\frac{d_p}{2}} x^2 \, dx \rightarrow \sigma_i = \frac{d_p}{\sqrt{12}}. \tag{2.7}$$

Assuming planar detectors 1 and 2, and a perpendicular particle track, the two limiting cases, where the intrinsic resolution of either detector is perfect ($\sigma_1 = 0$ or $\sigma_2 = 0$, respectively), can be expressed as:

$$\left.\frac{\sigma_g}{\sigma_1}\right|_{\sigma_2=0} = \frac{r_2}{r_2 - r_1} \quad \text{and} \quad \left.\frac{\sigma_g}{\sigma_2}\right|_{\sigma_1=0} = \frac{r_1}{r_2 - r_1}, \tag{2.8}$$

such that the geometrical contribution σ_{DCA}^g to the pointing resolution yields [59, 63]:

$$\sigma_{DCA}^g = \sqrt{\left(\frac{r_2}{r_2 - r_1} \sigma_1\right)^2 + \left(\frac{r_1}{r_2 - r_1} \sigma_2\right)^2} \tag{2.9}$$

Combining the contribution from multiple scattering (originating in the beam pipe or innermost layer) with the geometrical contribution by summing them in quadrature, the total impact parameter resolution is given by:

$$\sigma_{DCA} \sim \sigma_{DCA}^g \oplus \sigma_{DCA}^{MS}. \tag{2.10}$$

From this simplified model, the following design objectives therefore yield the best pointing resolution:

- Minimisation of the material budget X/X_0 to reduce the multiple scattering contribution (especially for the beam pipe and first detector layer).
- Smallest possible inner detector radius r_1, to decrease the weighting term of the intrinsic detector resolution in (2.9).
- Smallest intrinsic detector resolution σ_i (small pixel pitch d_p for a binary detector), especially for the innermost detector layer (where the resolution term is weighted by the radius r_2).
- Large 'lever arm' $(r_2 - r_1)$, reducing the intrinsic detector resolution scaling factor.

This leads to a complex optimisation problem. For semiconductor pixel tracking detectors, factors such as power density (and hence temperature), readout bandwidth, cost (particularly for large-area outer layers), redundancy, and fabrication and technology constraints all need to be considered. For the ALICE ITS2, 3 IB tracking layers and 4 OB tracking layers were chosen as discussed above. A more exhaustive discussion on impact parameter resolution, including multiple layers, the magnetic field, and track geometry, can be found in [59, 64, 65]. The impact parameter resolution, as the dispersion of DCA for tracks originating at the PV – in both the transverse plane ('DCAxy') and longitudinal plane ('DCAz') – will be used going forward (as also defined in [65]).

2.4 ITS3 upgrade

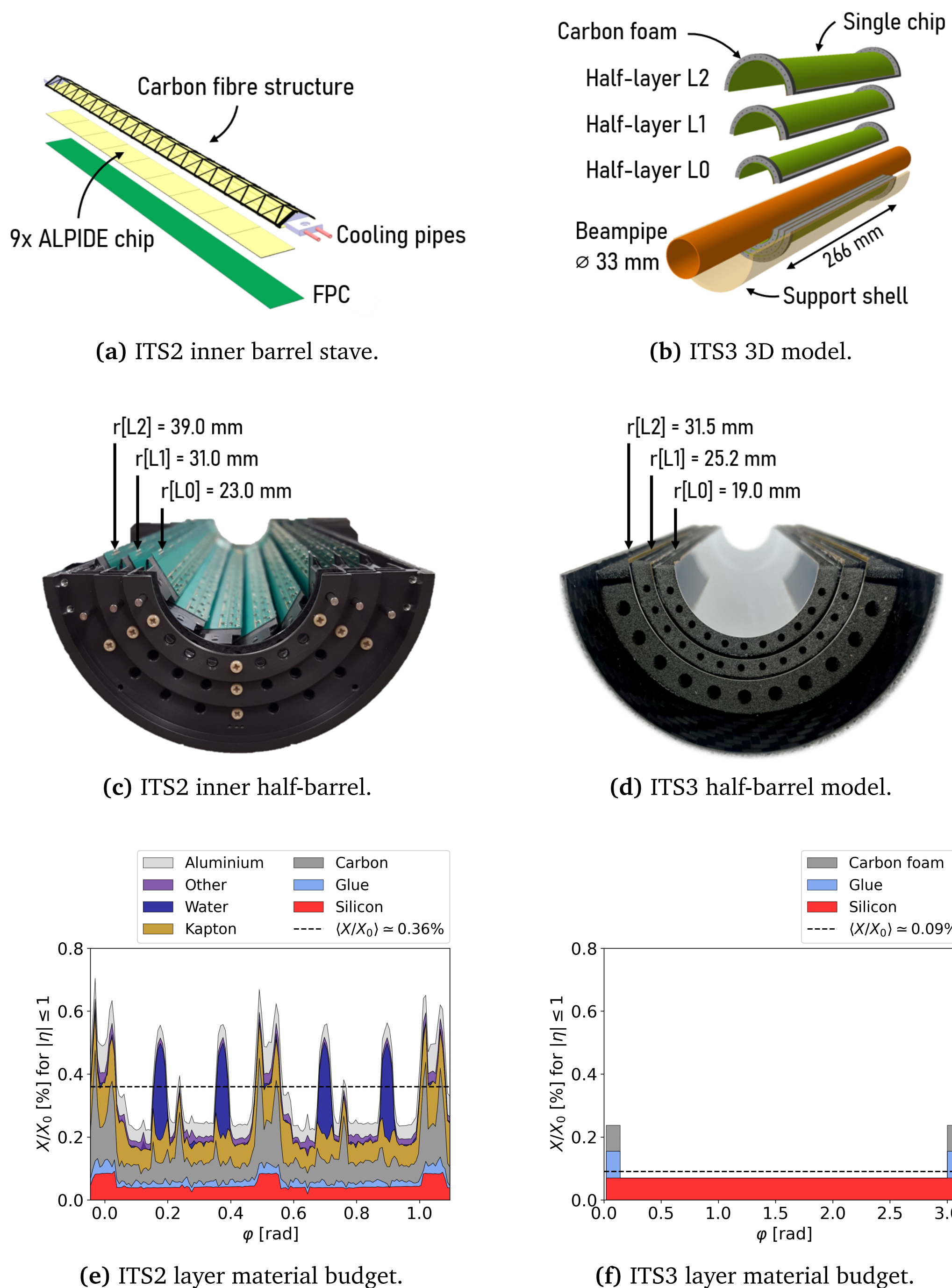
To further improve on the impact parameter resolution (and tracking efficiency) at low transverse momenta $p_T \lesssim 10$ GeV/c, the Inner Tracking System 3 (ITS3) upgrade is being developed. The inner barrel (i.e. the innermost three layers) of the currently installed ITS2 will be replaced by the ITS3. In the remainder of this thesis, the following naming convention will be used: ITS refers to the inner tracking system as

a whole, ITS2 to the currently installed configuration (wherever this differentiation is needed), and ITS3 to the standalone inner barrel upgrade. ITS, in this text, never refers to the inner tracking system installed for the initial ALICE Run 1 configuration. The simplified schematic drawing of the ITS3 is shown in Figure 2.8b.

The ITS3 vertexing and tracking detector consists of two half barrels, each with three layers L0, L1, and L2 of monolithic, cylindrically bent, wafer-scale pixel sensor planes (green). They are held in place by carbon foam structures: two 'half rings' at each end, and two 'longerons' along the beam axis. The sensors are air-cooled. Below a thickness of about 50 μm, silicon sensors become flexible, and required bending radii were shown to be easily achieved [66], while the devices stay fully operational [67]. Wire bonding on a curved surface was demonstrated successfully [68].

The ITS3 sensors will be manufactured in 65 nm CMOS TPSCo (Tower Partners Semiconductor Co.) imaging technology, employing a so-called stitching fabrication technique (see Section 3.4). This approach enables the production of the large monolithic sensor planes required. Compared to the ITS2 inner barrel, the ITS3 achieves a drastic reduction of about a factor 4–5 in material budget from $\langle X/X_0 \rangle_{L_i} \simeq$ 0.36% to $\langle X/X_0 \rangle_{L_i} \simeq 0.09\%$ per layer (sensor-only contribution: $min(X/X_0)_{L_i} \simeq$ 0.07%). This is illustrated in Figure 2.8, showing the components, design, layout and material budget contribution for the ITS2 inner barrel (a, c, e) and ITS3 (b, d, f), respectively. Removal of (most) mechanical support, the cooling plate, and Flexible Printed Circuit (FPC – integrated on-chip for the ITS3 sensors) allows for the ultra-light design of the ITS3. The general parameters of the ITS3 are summarised in Table 2.1.

An exploded view of one innermost half layer, L0, of ITS3 is shown in Figure 2.9, illustrating the interconnection, mounting, and cooling strategy. The sensor is electrically interconnected via wirebonds to two FPCs: one provides power only, and the second provides both power and data communication. The sensor is mechanically glued and held in place by carbon foam structures. The half-ring support structure doubles as a heat radiator, since the power density in the chip periphery, where the data links are located, can reach up to 1,000 mW cm^{-2}. For air cooling to be sufficient, the average power density in the sensitive area of the chip is limited to approximately



(a) ITS2 inner barrel stave.

(b) ITS3 3D model.

(c) ITS2 inner half-barrel.

(d) ITS3 half-barrel model.

(e) ITS2 layer material budget.

(f) ITS3 layer material budget.

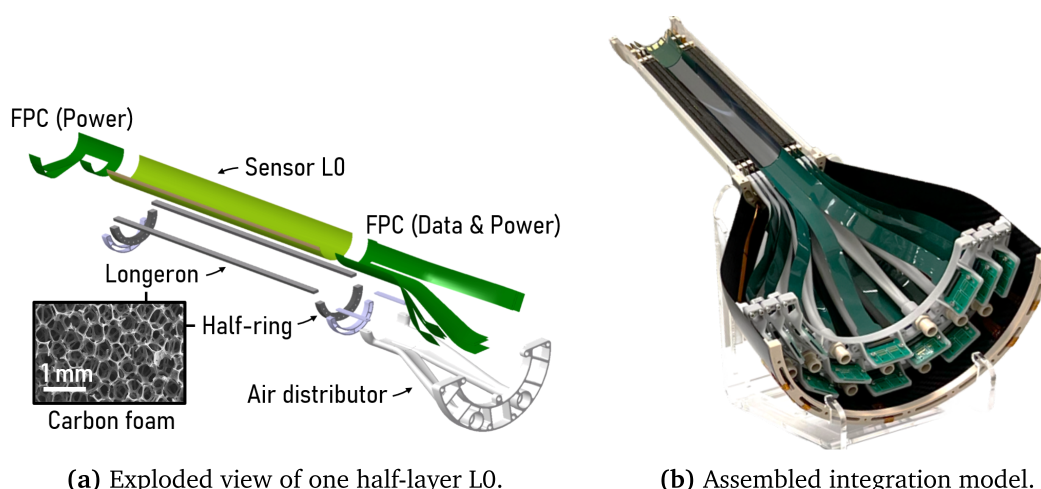
Figure 2.8: (a) ITS2 single inner barrel stave with 9 ALPIDE pixel chips, carbon fibre support frame, cooling plate with pipes, and Flexible Printed Circuit board (FPC) [54]. (b) Simplified ITS3 detector layout [6]. (c) One half of the currently installed ITS2 inner barrel with layer mean radii indicated. The outer beam pipe diameter is 36.0 mm. (d) One half of the ITS3 inner barrel as engineering model (dummy silicon layers). The cylindrical support structure is made from carbon fibre. (e) ITS2 IB material budget. Adapted from [52]. (f) ITS3 L0 material budget (silicon here includes the on-chip copper metal stack, see Section 3.5.2, and is therefore slightly larger than in (e)). Adapted from [6].

40 mW cm^{-2}. To verify the thermal performance, a dedicated test setup was built using copper serpentines embedded in a polyimide–silicon sheet stack to simulate the ITS3 layer power dissipation. At the target power levels in both the periphery and the sensitive area, the resulting thermal gradient along the chip remained below 5 K. A temperature increase of below 5 K above the air inlet temperature was achieved at an 8 m/s freestream velocity between layers [69]. Aeroelastic performance studies showed a maximum radial sensor displacement induced by air flow at 8 m/s remained below ±0.5 µm peak-to-peak (integrated $RMS_{flow} < 0.4$ µm), which is well within the specification envelope of $RMS_{total} < 2.0$ µm for the total short term displacement [6, 70]. The Coefficients of Thermal Expansion (CTE) of the used materials are closely

Table 2.1: General ITS3 parameters [6].

Beampipe inner/outer radius (mm)	16.0 / 16.5		
ITS3 parameters	**Layer 0**	**Layer 1**	**Layer 2**
Radial position (mm)	19.0	25.2	31.5
Length (sensitive area) (mm)	260	260	260
Pseudo-rapidity coverage[a]	±2.5	±2.3	±2.0
Active area (cm^2)	305	407	507
Pixel sensors dimensions (mm^2)	266 × 58.7	266 × 78.3	266 × 97.8
Number of pixel sensors / layer	2		
Equatorial gap (top–bottom) (mm)	1.0		
Material budget (% $[X/X_0]$ / layer)	0.086		
Silicon thickness (µm / layer)	≤ 50		
Pixel size (µm^2)	20.8 × 22.8		
Single point resolution (µm)	$\lesssim 5$		
Fractional sensitive sensor area (%)	> 92		
Detection efficiency (%)	> 99		
Fake-hit rate (pixel^{-1} s^{-1})	< 0.1		
Fake-hit occupancy (pixel^{-1} frame^{-1})	$< 10^{-6}$ (@ 10 µs frame duration)		
Frame duration (programmable) (µs)	2–10		
Power density (mW/cm^2)	<40/<1000 (sensitive area/periphery)		
NIEL (1 MeV n_{eq} cm^{-2})	10^{13}		
TID (kGy)	10		
Target operating temperature (°C)	15–30		
Pb-Pb interaction rate (kHz)	50/164 (average/peak w. 2x safety)		
Total particle flux[b] (MHz cm^{-2})	3.20+2.55 (QED electrons + Hadronic)		

[a] For tracks originating from a collision at the nominal interaction point ($z = 0$).
[b] At 164 kHz interaction rate, Layer 0, $z = 0$.



(a) Exploded view of one half-layer L0.

(b) Assembled integration model.

Figure 2.9: (a) Exploded view of one half-layer L0 with annotated components [6]. (b) Assembled integration model with three dummy silicon layers, power and data FPCs, air distributor, and cylindrical support structure [71].

matched to limit stress induced by thermo-elastic deformation, and thermal cycling of dedicated model assemblies was successfully performed between 10–40°C [6]. Discussion of the sensor technology follows in Sections 3.3–3.6. A fully assembled ITS3 integration model (one half-barrel) is depicted in Figure 2.9b.

The ITS3 design leads to an improved impact parameter resolution by:

- reduction of the material budget by a factor ~ 5 compared to the ITS2 (IB),
- reduction of the innermost layer radius from 23.0 mm $\rightarrow$ 19.0 mm,
- reduction of the wall thickness of the beryllium beam pipe from 0.8 mm $\rightarrow$ 0.5 mm.

2.4.1 Physics performance simulation

Physics performance studies of the ITS are based on a full Monte Carlo (MC) and track-reconstruction algorithm, developed within the ALICE O^2 framework for Run 3 [6, 72, 73]. The ITS3 geometry and material (including carbon foam, glue, sensor dead zones, and support structure) are implemented. The simulation and reconstruction sequence consists of four main steps:

1. Generation of particles by simulation of physics events using `PYTHIA 8` [74] for *pp* collisions and `HIJING` [75] for heavy-ion collisions.

2. Particle transport through the detector using `GEANT 4` [76] and hit production.

3. Digitisation by simulation of the detector response from the particle energy loss in the active material.

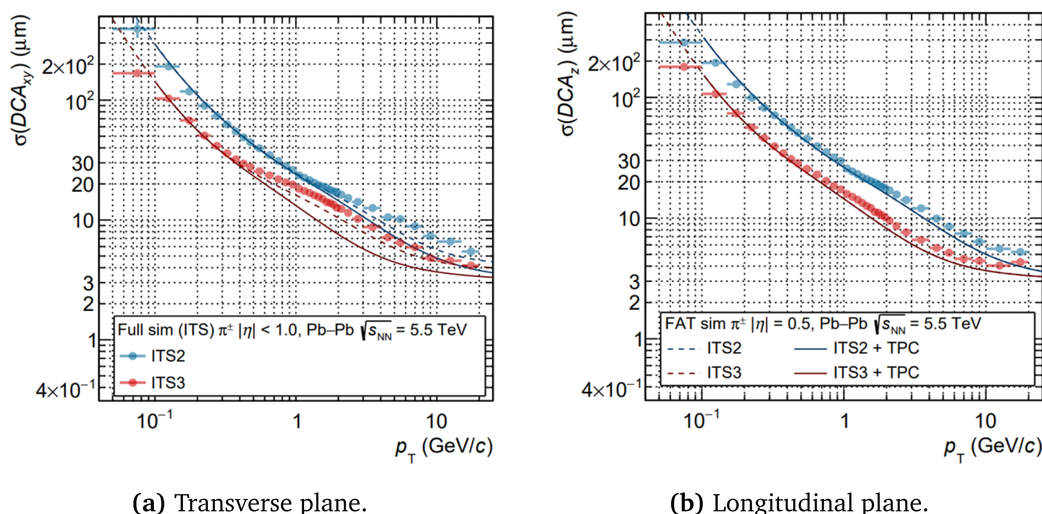
4. Track-reconstruction (see below).

To assess the ITS standalone tracking performance, the existing ITS2 software was adapted to include the ITS3 inner barrel [77]. It utilises all seven ITS tracking layers. After seed finding with the three ITS3 layers, candidates are identified by possible connections across layers, and tracks are fitted using a Kalman filter (including MS and energy losses), with the best candidates selected by χ^2 based evaluation. The same framework and simulation sequence are used in the performance studies discussed in Chapter 8.

In addition to the ITS standalone tracking performance, the overall ALICE tracking performance will improve. The secondary vertex reconstruction capability is well measured by the track impact-parameter resolution, as discussed above. For transverse momenta below $p_T \lesssim 10$ GeV/c, a factor 2 improvement in both transverse (DCAxy) and longitudinal (DCAz) impact parameter resolution is expected, as shown from simulations in Figure 2.10a and Figure 2.10b, respectively [6]. Only pion tracks with $|\eta| < 1$ and hits in all ITS layers were selected. At high p_T, the impact parameter resolution is expected to approach the intrinsic sensor resolution of $O(5\ \mu\text{m})$. The difference in impact parameter resolution between transverse and longitudinal simulations is attributed to the transverse-momentum resolution of the ITS, which is significantly improved by extrapolating the tracks with the TPC. A Fast Analytic Tool (FAT), based on a simplified detector model, was used to estimate this effect. For $p_T > 0.5$ GeV/c an improvement in DCA resolution in the transverse plane is visible due to the improved p_T resolution.

From the full simulation, the ITS3 achieves a track reconstruction efficiency above 90% for transverse momenta down to 100 MeV/c, as shown in Figure 2.11. At lower momenta, an improvement of the track-reconstruction efficiency of approximately 30% is expected with the ITS3 over the current ITS2 configuration. A slightly higher

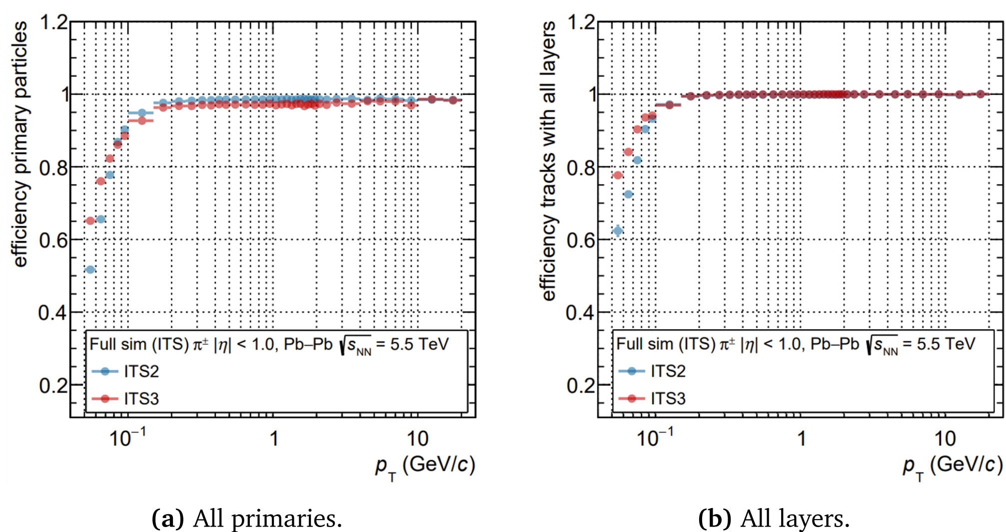
fake rate, where a track is reconstructed with a hit not associated with the charged particle producing the track, is expected for the ITS3. When requiring hits in all ITS tracking layers (see Figure 2.11b), an even higher track reconstruction efficiency of 95% is achieved down to transverse momenta of 100 MeV/c.



(a) Transverse plane.

(b) Longitudinal plane.

Figure 2.10: Comparison of the ITS performance with the currently installed ITS2 and after the ITS3 upgrade [6]. (a) Impact parameter resolution in the transverse plane. (b) Impact parameter resolution in the longitudinal plane.



(a) All primaries.

(b) All layers.

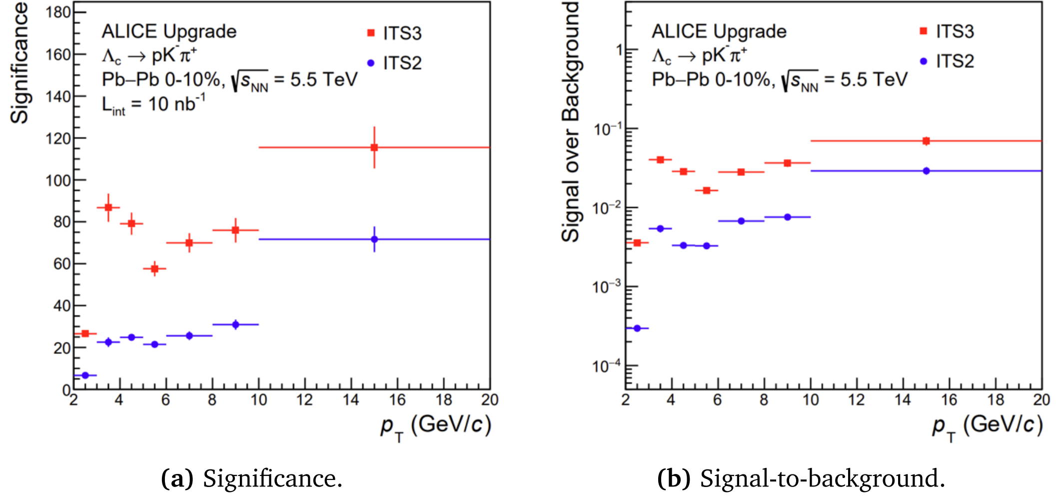
Figure 2.11: Comparison of the ITS tracking efficiency with the currently installed ITS2 and after the ITS3 upgrade [6]. (a) For all primary charged pions. (b) For tracks with hits in all 7 ITS tracking layers.

2.4.2 Λ_c^+ measurement

One benchmark channel for ITS performance is the measurement of Λ_c^+ baryon production. The Λ_c^+ is the lightest charmed baryon, with a mass of $m(\Lambda_c^+) = 2286.46 \pm 0.14\,\text{MeV}$, proper decay length $c\tau \sim 60.75\,\mu\text{m}$, and quark content $\Lambda_c^+ = udc$ [33]. Measurements of the production yield and flow of charm (and beauty) baryons are of interest for the thermalisation and hadronisation mechanisms of c and b quarks in the QCD medium. If heavy quarks thermalise and hadronise via recombination with light-flavour quarks in the QGP or its phase boundary, an enhancement in charm (and beauty) baryon production is expected in the momentum region below about $10\,\text{GeV}/c$, when comparing Pb-Pb and pp collisions. Such enhancement of the baryon-to-meson ratio was already measured in the light-flavour sector. However, precise measurements in the charm sector are still lacking and would provide crucial insights into charm-quark thermalisation and hadronisation in the QGP, as well as the relative roles of recombination and radial flow [21]. First measurements of the Λ_c^+/D^0 ratio at $p_T < 10\,\text{GeV}/c$ in Pb-Pb (at $\sqrt{s_{NN}} = 5.02$ TeV) and Au-Au (at $\sqrt{s_{NN}} = 200$ GeV) collisions [78, 79] indicate an enhancement with no modification for higher p_T. However, the statistical precision of these measurements is insufficient to draw firm conclusions [80].

The most convenient decay channels are $\Lambda_c^+ \to pK^-\pi^+$, which suffers from a large three-prong combinatorial background, and $\Lambda_c^+ \to pK_S^0$ which does not allow for precise decay vertex determination given the long decay length of the neutral kaon. Recently, machine learning techniques have been employed to exploit the pK_S^0 channel as well [81]. Focusing on the three-prong $\Lambda_c^+ \to pK^-\pi^+$ decay, outstanding impact parameter resolution is required for successful secondary vertex reconstruction, given the short decay length ($c\tau \sim \gamma\beta c\tau$ for $|\mathbf{p}(\Lambda_c^+)| \sim m(\Lambda_c^+)$). This makes it an ideal benchmark for evaluating the performance of the ITS3.

Topological cuts on the secondary vertex allow stronger rejection of combinatorial backgrounds and larger efficiency for signal selection. The performance with the ITS3 compared to the ITS2 was simulated as outlined in [21, 54], and illustrated in Figure 2.12a and Figure 2.12b showing the statistical significance $S/\sqrt{S+B}$ and



(a) Significance. **(b)** Signal-to-background.

Figure 2.12: Statistical significance (a) and signal-to-background ratio (b) as a function of p_T for $\Lambda_c^+ \rightarrow pK^-\pi^+$ in central Pb-Pb collisions, comparing the performance of the ITS2 and after the ITS3 upgrade [21].

signal-to-background ratio S/B, respectively. An improvement of approximately a factor of four in statistical significance and a factor of ten in S/B is observed. Especially at low transverse momenta, this improvement will be crucial for precise Λ_c^+ production yield measurements and for determining the total $c\bar{c}$ cross-section, which is a key input to model charmonium re-generation in the QGP [21].

In Section 8.5.3 of this thesis, the Λ_c^+ three-body decay to $pK^-\pi^+$ is used to study the performance loss in tracking efficiency, if parts of the ITS3 sensor planes malfunction. To isolate the p_T-dependent tracking loss of the daughter particles, a simplified Monte Carlo (MC) simulation chain is used. The parent Λ_c^+ momentum is sampled from a `FONLL` D^0 prediction, scaled by the measured Λ_c^+/D^0 ratio [82] for prompt production in pp collisions at $\sqrt{s} = 5.02$ TeV. The resulting Λ_c^+ p_T-spectrum is shown in Figure 8.23a. The decay chain is simulated with `PYTHIA` (see also Section 8.5.3), exclusively enabling the decay modes given in Table 2.2 – with the corresponding branching ratios as indicated. The three-body decay may occur via resonant intermediate two-body states, which are assumed to decay promptly in this study.

The kinematic phase space of a three-body decay can be illustrated with a Dalitz plot as shown in Figure 2.13. The resonant structures appear over the non-resonant phase space, which is otherwise flat, with all points equally probable within the allowed

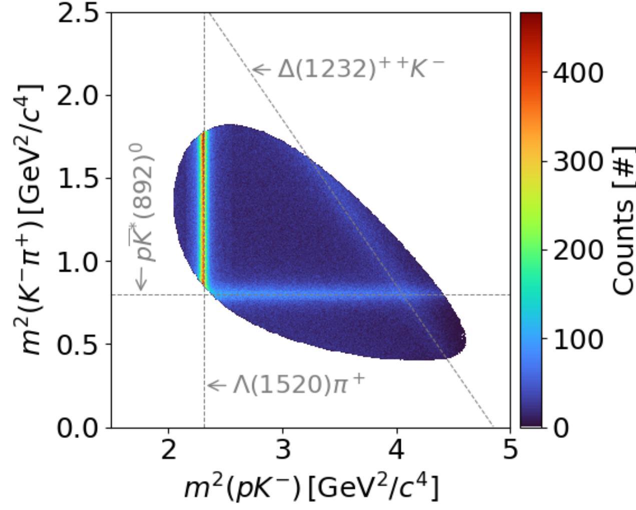


Figure 2.13: Dalitz-plot illustrating the Λ_c^+ kinematic phase space for $pK^-\pi^+$ final states as used in the simulation. On the smooth, non-resonant phase space, the resonant decays are visible and annotated.

kinematic boundaries. The decay rate of a particle (Λ_c^+) with invariant mass M is given in terms of the Lorentz-invariant, spin-state-averaged matrix element $\mathscr{M}$ by [33]:

$$d\Gamma = \frac{1}{(2\pi)^3} \frac{1}{32M^3} \overline{|\mathscr{M}|^2} \, dm_{12}^2 \, dm_{23}^2 \,, \tag{2.11}$$

where the two-body invariant masses are defined as $m_{ij}^2 = p_{ij}^2$ and $p_{ij} = p_i + p_j$. Here, $m_{12}^2 = m^2(pK^-)$ and $m_{23}^2 = m^2(K^-\pi^+)$, which correspond to the Dalitz plot axes.

From momentum–energy conservation $m_{12}^2 + m_{23}^2 + m_{13}^2 = M^2 + m_1^2 + m_2^2 + m_3^2$ follows, and any two m_{ij}^2 fully specify the three-body decay kinematics. A uniform event distribution in the Dalitz plot corresponds to a constant $\overline{|\mathscr{M}|^2}$, while deviations from uniformity directly reflect the dynamics of the decay through $\overline{|\mathscr{M}|^2}$. The quantities $m_{ij}^2 = (p_i + p_j)$ are readily measurable in experiments and allow for the determination

Table 2.2: Decay modes of $\Lambda_c^+ \to pK^-\pi^+$ used in the simulation. From [83].

Intermediate state	Γ_i/Γ (%)
$\overline{K}^*(892)^0 \to K^-\pi^+$	1.96 ± 0.35
$\Delta(1232)^{++} \to p\,\pi^+$	1.08 ± 0.31
$\Lambda(1520) \to p\,K^-$	2.2 ± 0.8
non-resonant phase space	3.5 ± 0.4
Inclusive $\mathscr{B}(\Lambda_c^+ \to pK^-\pi^+)$	6.28 ± 0.32

of resonant structures and amplitude measurements. A recent $\Lambda_c^+ \to pK^-\pi^+$ decay study at LHCb in [84], provides an example of such an analysis, presenting a Dalitz plot representation and amplitude measurement of the resonant contributions.

3

Monolithic Active Pixel Sensors, MOSS, and ITS3 sensors

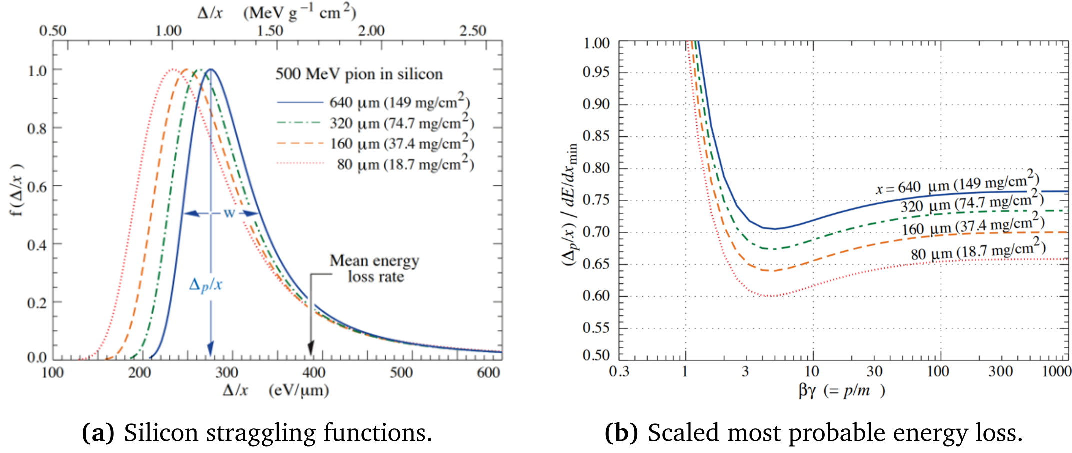
In this chapter, after an introduction to the interaction of charged particles with matter and silicon, the working principle and fabrication of semiconductor-based monolithic active pixel sensors are discussed. The prototype MOnolithic Stitched Sensor (MOSS) is described in detail, followed by a discussion of the layout and geometry of the final ITS3 sensor layers.

3.1 Interaction of particles with matter

Particle detection requires interaction of particles with matter. For silicon detectors, the most relevant mechanism is ionisation of a medium by a traversing charged particle. The mean rate of energy loss (or stopping power) of a moderately relativistic heavy charged particle is described by the Bethe equation [33]:

$$\left\langle -\frac{dE}{dx} \right\rangle = Kz^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2m_e c^2 \beta^2 \gamma^2 W_{\text{max}}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right]. \tag{3.1}$$

Here, the stopping power is expressed in units of $\text{MeV}\,\text{g}^{-1}\text{cm}^{-2}$, and the equation is valid to within a few percent for $0.1 \lesssim \beta\gamma\,(= p/Mc) \lesssim 1{,}000$. The coefficient $K\,[\text{MeV}\,\text{mol}^{-1}\text{cm}^2] = 4\pi N_A r_e^2 m_e c^2$, and z is the charge number of the incident particle of mass M. The atomic number and atomic mass of the absorber material are Z and



(a) Silicon straggling functions.

(b) Scaled most probable energy loss.

Figure 3.1: (a) Straggling functions for 500 MeV pions in silicon, scaled to unity at the Most Probable Value (MPV) Δ_p/x. (b) Most probable energy loss, in the standard Bethe representation, when scaled to the mean loss of a minimum ionising particle (388 eV/µm). Figures from [33].

A, respectively. The maximum possible energy transfer to an e^- in a single collision is denoted by $W_{\max}$. The mean excitation energy is given by I, and $\delta(\beta\gamma)$ accounts for the density effect correction to the ionisation energy loss.

The energy loss via ionisation in a medium is inherently stochastic, given that each elastic collision can be treated as an independent event. Therefore, a Probability Density Function (PDF) is used to describe the most probable energy loss Δ_p/x and its fluctuations, known as straggling functions. For very thin (e.g. $\lesssim 300\,\mu\text{m}$ for silicon) absorbers, such as the silicon devices discussed in this work, Bichsel functions provide a very good description [85]. In Figure 3.1, such straggling functions are illustrated for multiple silicon thicknesses and 500 MeV incident pions. Therefore, in a sensitive silicon layer with $O(10\,\mu\text{m})$ thickness and a mean electron-hole (eh) pair creation energy of ~ 3.65 eV, about 50 eh/μm are generated by a Minimum Ionising Particle (MIP), as discussed below[1].

For photons (i.e. electromagnetic radiation), the most relevant energy loss mechanisms are the photoelectric effect, Compton scattering, and pair production [87, 88]. The contributions of these processes to the total photon cross section in silicon are shown in Appendix A.1 as a function of photon energy, based on data from [89].

[1]An extension to the model illustrating the energy loss in e.g. 10 µm thin silicon can be found in [86].

3.2 Silicon as detector material

Semiconductors, and specifically silicon, are commonly used as a material for particle detectors in high-energy physics. Silicon crystallises in a diamond lattice structure, with each silicon atom covalently bonded to four nearest neighbours [59, 90]. Given the dense periodic lattice arrangement, energy levels of individual atoms are split by the influence of many neighbouring atoms, and grouped into energy bands.

Energy bands are separated by a band gap, with the highest energy bands the Valence Band (VB) and Conduction Band (CB) separated by the bandgap energy E_g. For silicon $E_g \sim 1.12$ eV at room temperature (300 K) as illustrated in Figure 3.2 (left). Within a band, energy levels are so closely spaced that transitions to unoccupied levels occur easily. Conduction, therefore, depends on the band occupation. Owing to the small energy band-gap, (weaker bonds between neighbour atoms than insulators), electrons can rise to the conduction band through thermal excitation or external electric fields. A broken bond leaves a mobile hole h in the VB and a mobile electron e in the CB.

In thermal equilibrium, for an intrinsic (undoped) silicon semiconductor, the concentration of holes in the VB and electrons in the CB is equal, and described by the charge carrier density $n_i = n_h = n_e$ (or, as generation-recombination rates are equal, $n_i^2 = n_h \cdot n_p = const.$). The intrinsic carrier concentration is highly temperature dependent as $n_i \propto T^{3/2} \exp(-E_g/2k_B T)$, with a value for silicon at 300 K of $n_i \sim 1.01 \cdot 10^{10}\,\mathrm{cm}^{-3}$ [91]. With carrier mobilities $\mu_e = 1,350\ \mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ and $\mu_h = 480\ \mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$, the resistivity is given by [92, 93]:

$$\rho\,[\Omega\,\mathrm{m}] = \frac{1}{e\,(n_e\mu_e + n_h\mu_h)} = \frac{1}{e\,n_i\,(\mu_e + \mu_h)}\,. \tag{3.2}$$

The average energy w_i required for the generation of an eh-pair at 300 K (this value is also temperature dependent), such as by energy loss of an incident particle, is $\langle w_i \rangle = 3.65$ eV [94]. This value is higher than the band gap energy E_g, as part of the deposited energy is lost to lattice phonon excitations [59].

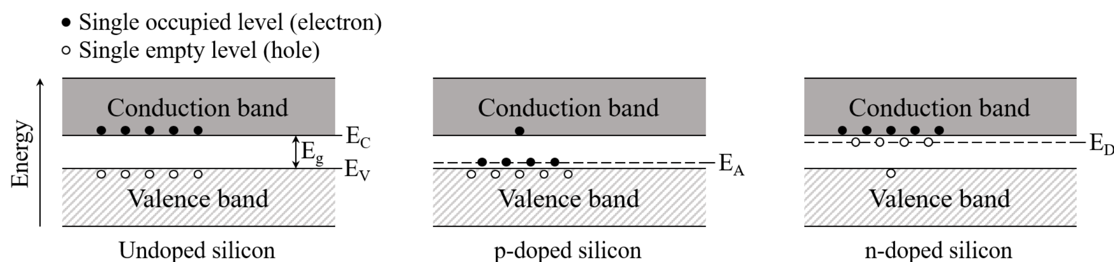


Figure 3.2: Band-model for undoped (left), p-doped (centre), and n-doped (right) silicon. See text.

With controlled introduction of impurities, the conduction properties of silicon can be altered. A pentavalent element (e.g. P), called a donor, placed in tetravalent silicon creates an excess of conduction electrons. This is referred to as n-doping. Doping with a trivalent element (e.g. B), called an acceptor, creates an excess of holes, and is referred to as p-doping.

The alteration in energy band structure is illustrated in Figure 3.2 (centre, right), showing the introduced acceptor level E_A (p-doped), and donor level E_D (n-doped). The mass-action law $n_i^2 = n_h \cdot n_p$ is still valid for doped (extrinsic) semiconductors, as the increase in one type of carrier density is compensated by a corresponding decrease in the other. If impurities are uniformly distributed, the net charge density in every volume element is zero. For p-doped silicon, holes become the majority charge carrier, while in n-doped silicon, electrons are the majority. Other impurities, such as radiation-induced defects, can create additional energy levels that act as generation-recombination centres and lead to an increase in detector leakage current[2] [88, 95, 96].

3.2.1 p-n junction and charge carrier transport

To create a functioning semiconductor detector, the interface between n-doped and p-doped silicon is exploited. A p-n (or pn) junction is schematically shown in Figure 3.3. At the p-n interface, recombination of the two carrier types – electrons from the n-layer and holes from the p-layer – occurs, creating a charge carrier-free 'space charge region', commonly called the 'depletion zone'. While there are no free charge

[2]Defined as the steady state current of a reverse biased p-n junction – see next section – which is also strongly temperature dependent $J \propto T^2 e^{-E_g(T)/2k_B T}$ [88].

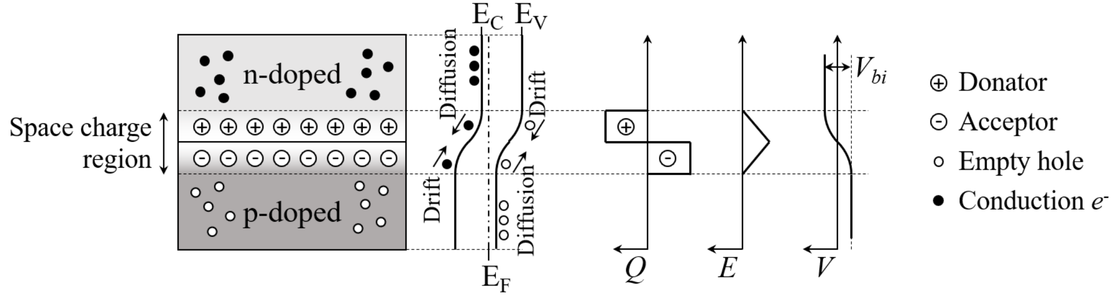


Figure 3.3: Schematic illustration of a p-n junction with band model, charge, electric field, and potential along the device.

carriers, the p-layer and n-layer feature a negative and positive space-charge density $\rho(x)$, which only depends on the respective doping concentration. The opposite space charge in the p and n layer leads to the formation of an intrinsic electrical field, with drift and diffusion currents I_{drift} and I_{diff} (as indicated in Figure 3.3 in equilibrium (drift and diffusion are discussed further below). The resulting built-in potential V_{bi} (from derivation via Poisson's equation) is given by [59]:

$$V_{bi} = \frac{k_B T}{e} \ln\left(\frac{N_A N_D}{n_i^2}\right) \overset{Si}{\sim} 0.4 - 0.8\ V\ , \tag{3.3}$$

with N_A, N_D the acceptor and donor concentrations in the p and n regions, respectively.

If an external voltage V_R is applied in 'reverse bias' configuration (with the electric field directed from n+ to p–), the depletion region free of charge carriers is enlarged. For a planar junction, the total space charge width is computed as [92]:

$$W = \left[\frac{2\,\epsilon_s}{e}\left(\frac{N_A + N_D}{N_A N_D}\right)(V_{bi} + V_R)\right]^{1/2}\ , \tag{3.4}$$

where ϵ_s is the silicon permittivity. The junction capacitance for area A, in the planar case, is written (analogous to a parallel plate capacitor) as $C = \epsilon_s \frac{A}{W}$. Therefore, an increase in reverse bias decreases the device capacitance.

Once an electron-hole pair is created in silicon, the charge carriers migrate through two transport mechanisms: drift – movement due to electric fields, and diffusion

– flow of charge due to density gradients. The moving charges generate currents, described by the current density equation $J = J_{\text{drift}} + J_{\text{diff}}$. Propagating charge carriers are subject to scattering or collision, dominated by lattice (phonon) scattering and (ionised, defect) impurity scattering. The carrier mobility, μ_e for electrons and μ_h for holes, is used to describe these effects, and is strongly dependent on temperature and impurity concentration [92, 97]. Drift, as the induced movement of charge carriers by an electric field E, is characterised by the average drift velocity

$$v_{\text{drift}\,\{e,h\}} = \mu_{\{e,h\}}\,E \ , \tag{3.5}$$

and the total drift current density as the sum of the electron and hole drift current densities (and related to the resistivity ρ from Equation (3.2)):

$$J_{\text{drift}} = J_{\text{drift},\,e} + J_{\text{drift},\,h} = e\,(n_e\mu_e + n_h\mu_h)\,E = \rho^{-1}E\ . \tag{3.6}$$

Diffusion, conversely, is governed by the spatial variation of carrier concentration in the material and thermal random walk. A diffusion current flows from the high concentration region to the low concentration region along the direction of the e, h density gradients ∇n_e, ∇n_h [92, 98]:

$$J_{\text{diff}} = J_{\text{diff},\,e} + J_{\text{diff},\,h} = eD_e\nabla n_e - eD_h\nabla n_h \tag{3.7}$$

The diffusion constants $D_{\{e,h\}}$ are related to the charge carrier mobilities by the Einstein relation $D_{\{e,h\}} = \frac{k_B T}{e}\,\mu_{\{e,h\}}$.

Another ingredient is the average carrier lifetime $\langle\tau\rangle$ of a generated eh-pair. Annihilation or recombination of an eh-pair can occur at trapping or recombination centres, such as impurities or defects, which are capable of capturing the carriers. To ensure efficient charge collection, the collection time in a detector must be much shorter than $\langle\tau\rangle$, which is generally the case for the sensitive volumes of the devices discussed below[3] [95]. The dependency on the doping concentration and general statistical treatment of recombination are found at [99, 100].

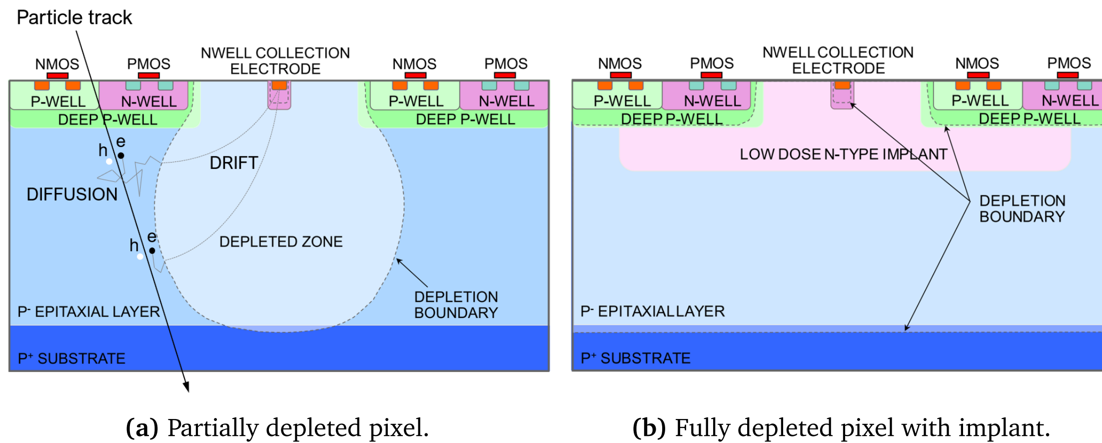
[3]For high doping concentrations, such as in the collection electrode and substrate of the sensors discussed below, the recombination contribution becomes relevant.

Qualitatively, a silicon tracking and vertexing detector such as that used for the ALICE ITS should meet several key requirements. It should be thin, to minimise the material budget, and provide high intrinsic spatial resolution. Charge collection should be fast, relying primarily on drift (rather than diffusion) and a large depletion volume. The signal generated must be significantly larger than intrinsic thermal excitations, which requires low leakage current – achieved through low operating temperatures and/or high-resistivity silicon. The device capacitance should be small, enabling low power consumption, a large charge-to-capacitance ratio (Q/C), and a high signal-to-noise ratio. Finally, the detector must provide the required radiation hardness to withstand both non-ionising energy loss (NIEL) and total ionising dose (TID) effects.

3.3 Monolithic Active Pixel Sensors – MAPS

Monolithic Active Pixel Sensors (MAPS), as discussed in this text, are based on commercial CMOS (Complementary Metal-Oxide-Semiconductor) imaging technology. This technology allows for the integration of both the detection volume (based on a depleted p-n junction) and readout circuitry on the same piece of silicon – in contrast to the widely used hybrid pixel detectors (e.g. in ATLAS and CMS), where the sensor and readout elements are manufactured separately and then physically bonded. The manufacture of $\lesssim 50\,\mu$m thin (low material budget), low-power (air cooling is sufficient) pixel detectors required for the tracking and vertexing targets of the ALICE ITS becomes feasible with this approach.

The schematic cross-section for a single MAPS pixel is shown in Figure 3.4a as manufactured for the ALPIDE chip installed in the ITS2 [54, 55, 101, 102]. It is fabricated on a highly doped (p^+, $N_A \sim 10^{18}$ cm^{-3}) substrate, on which a high-resistivity (> 10 kΩcm) epitaxial layer (p^-, $N_A \sim 10^{13}$) is grown. A small-area (~ 2 µm diameter) n-doped collection electrode is embedded near the top of the epitaxial layer. By applying a reverse bias (< 10 V), a spherically shaped depletion zone forms, where charges, as generated by an incident particle track, are collected via drift. Outside the depletion zone, charge carrier transport relies solely on diffusion. A crucial development in this technology was the introduction of a deep p-well structure,



(a) Partially depleted pixel.

(b) Fully depleted pixel with implant.

Figure 3.4: Schematic cross section of a single MAPS pixel (a) partially depleted, with particle track and electron drift and diffusion indicated, and (b) fully depleted after introduction of a low-dose n-type implant below the collection electrode (this implant does not extend over the full pixel width but has a small 'gap' on each side). The epitaxial layer has a thickness of approximately 10 μm. Drawings not to scale. Discussion in text. Adapted from [4].

shielding the n-well of PMOS-transistors from competing with charge collection via the n-well collection electrode, and enabling full CMOS circuitry on-chip [103]. When an incident particle traverses the sensor, it liberates *eh*-pairs in the silicon. The liberated *eh*-pairs drift towards the collection electrodes within the electric field, and the movement of these charges induces a measurable signal on the electrodes (where the charge is ultimately collected), which is processed by the on-chip circuitry.

In the most recent iteration, the MAPS fabrication node was reduced from 180 nm (as used for ALPIDE) to 65 nm TPSCo (Tower Partners Semiconductor Company) imaging technology, as used for the sensors now discussed in this work. By introducing a low-dose n-type implant below the collection electrode, as shown in Figure 3.4b, the space charge region now immediately extends across the full epitaxial layer. This leads to substantially faster charge collection speeds, which can be further increased by applying reverse bias[4].

The smaller technology node allows for a smaller collection electrode (therefore smaller device capacitance and decreased power consumption), and a smaller pixel

[4]Faster charge collection solely by drift also increases the radiation hardness of a device, given the reduced trapping probability. Additionally, less charge sharing across pixels leads to an increased signal-to-noise ratio [59, 104, 105].

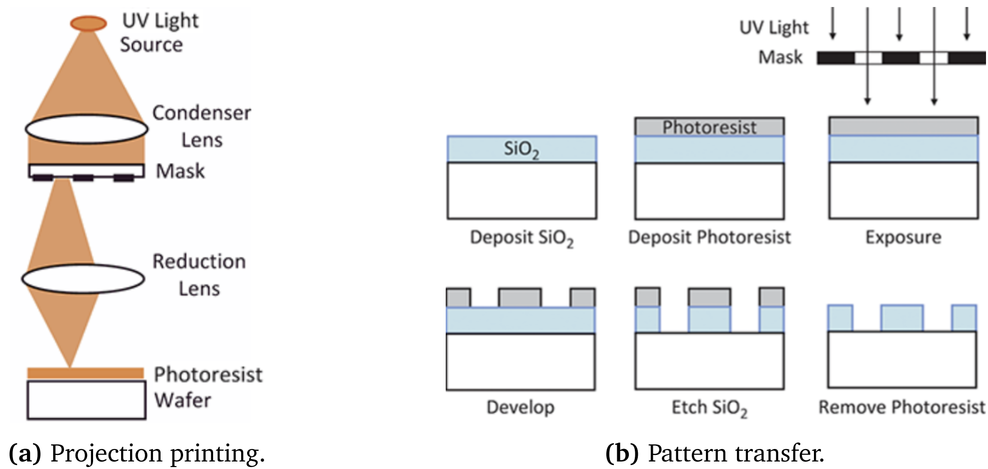
pitch. Furthermore, 300 mm diameter wafers, necessary for manufacturing the large-area sensor planes for the ITS3, as further discussed below, are available in this technology. An exhaustive program on test chips, manufactured in this technology [4, 104, 106], was conducted, validating its suitability in terms of charge collection efficiency, detection efficiency, radiation hardness, spatial resolution, fake-hit rate, and in-matrix power consumption required by the final ITS3 sensor [3, 107, 108].

3.4 Fabrication

CMOS device fabrication is a highly complex industrial process that undergoes continuous technological evolution. After a brief conceptual introduction to lithography, this section discusses two specific fabrication processes and device properties relevant to this work. For further details on CMOS fundamentals, fabrication techniques, and failure analysis, the reader is referred to [109–116].

3.4.1 Lithography

For the discussed technology node, photolithography is the primary tool to transfer a design pattern onto a silicon wafer, layer by layer. Conceptually, the process is depicted in Figure 3.5. From the chip design files, the stack of optical masks also referred to as design reticle(s), is manufactured. A widely used tool is 193 nm wavelength Deep-UltraViolet (DUV) immersion lithography, which enables sub-wavelength resolution. The field of view of such systems is usually a few centimetres on each side (corresponding to the effective reticle size), and the wafer is physically moved ('stepped') from one exposure field to the next. Individual regions of the reticle can be exposed at a time. On the wafer itself, depending on which layer is being manufactured, a photoresist is deposited before exposure with the UV lithographic system. After exposure, the photoresist is developed, and the pattern is etched (as an example in Figure 3.5b, into a dielectric layer of SiO_2), before the photoresist is removed.



(a) Projection printing.

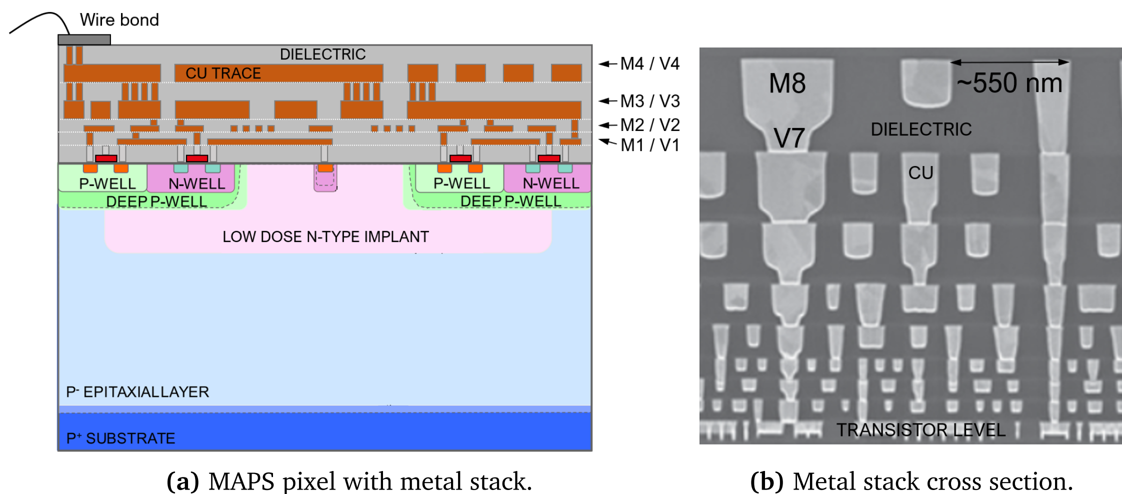
(b) Pattern transfer.

Figure 3.5: (a) Schematic illustration of DUV projection printing. (b) Lithographic pattern transfer. Adapted from [115].

3.4.2 Chip interconnect technology and metal stack

The on-chip metal stack is a crucial component, interconnecting the silicon device structures – such as transistors, (collection) diodes, resistors, and capacitors – to form a functional circuit. It is somewhat analogous to a multi-layer Printed Circuit Board (PCB) with metal layers and vertically connecting vias. An illustration of a MAPS pixel including the metals stack is shown in Figure 3.6a, with an electron microscopy cross-section image of a real device metal stack (not a MAPS) shown in Figure 3.6b. Cross-section images generated during failure analysis of the chips studied in this work, exposing their entire metal stack, cannot be disclosed here. With a smaller structure size (and therefore a smaller metal line cross-section area), such as in the 65 nm technology node and below, high conductivity is essential. Hence, copper (Cu) is used for metal interconnects instead of aluminium, which was used in the ALPIDE chip fabricated in 180 nm technology. Copper, however, cannot be effectively anisotropically etched chemically. To fabricate the copper-based metal stack, the so-called 'dual-damascene' process is employed (described below).

Metal layers (M#) are numbered, starting above the transistor level with 'M1' as they are built upwards. The lowest metal layers (or simply 'metals') form the local interconnects, realising short-distance circuit functions between transistor gates, sources, and drains. Given the short routing distance, a smaller line cross-section



(a) MAPS pixel with metal stack.

(b) Metal stack cross section.

Figure 3.6: (a) Illustration of a MAPS pixel with a metal stack on top. The actual devices have a larger number of metal layers. A wire bond pad and wire bond are indicated. Metals (M) are numbered from the transistor level upwards, with Vias (V) in between. (b) Cross-section electron microscopy image of an 8-layer metal stack in the Intel 32 nm node [117].

resulting in higher resistivity can be tolerated. Global, or long-distance interconnects, on the other hand, are routed on the uppermost metals covering large distances between circuit blocks. These upper layers carry the power grid that distributes supply voltage to all chip regions, as discussed further in Section 3.5.2 for the chip described in this work. Vertical connections between metal layers are termed Vias (V#), with the numbering scheme following the metals. The metal stack is embedded in dielectric, or Inter-Metal-Dielectric (IMD), which in its simplest form can be SiO_2. Modern devices, however, use so-called low-k dielectrics with a lower dielectric constant to reduce parasitic line-to-line capacitance, thereby improving the RC-limited switching speed of the device. It is important to note here that dielectrics can and do vary between metal layers, as the requirements for lower and higher metals differ.

One variation of the dual damascene process flow is illustrated in Figure 3.7. Dual, in this context, refers to the fact that vias and metal lines of one layer are manufactured in one iteration. The steps are as follows: (1.) On a previous metal layer (here, M1), two dielectric layers are deposited with etch stop layers (e.g. Si_3N_4, SiCN) on either side, and an additional hard mask layer on top. (2.) Using photolithography, the pattern which will become the metal 2 (M2) pattern is transferred and etched into the hard mask (stopped by the uppermost etch stop layer). (3.) A second round of

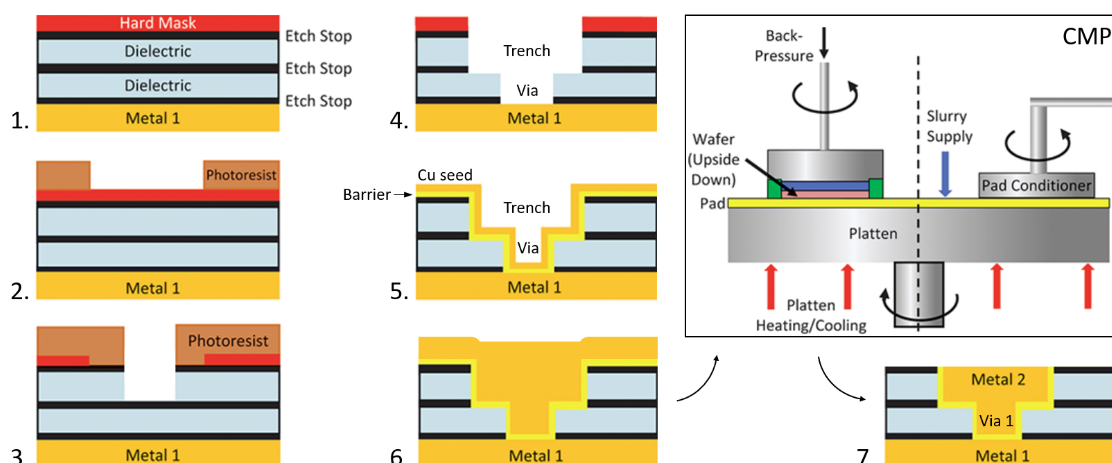


Figure 3.7: Dual damascene process steps and illustration of Chemical-Mechanical-Polishing (CMP). Adapted from [115].

photolithography (new application of photoresist layer) is performed, transferring the via 1 (V1) pattern into the upper dielectric layer, with the etching stopped by the middle etch stop. (4.) The photoresist is removed, and another anisotropic etching step is performed, such that the hard mask pattern (for M2) is transferred to the upper dielectric layer, and the previous V1 pattern (in the upper dielectric) is transferred to the lower dielectric. Both the via (for V1) and trench (for M2) are now formed, and the lower and middle etch stop layers are removed, exposing M1. (5.) A barrier layer (e.g. TaN/Ta) is deposited, serving as a Cu diffusion barrier, preventing copper from penetrating the (low-k) dielectric ('copper poisoning'). The TaN/Ta layer also increases the dielectric adhesion and Cu wettability. Next, a Cu seed layer is deposited, acting as a base for (6.) the electroplating of the bulk Cu. These are global processes; the entire wafer is now covered with a copper layer, and all V1 and M2 via/trench structures are filled. (7.) Chemical-Mechanical Polishing (CMP) is used to polish back the Cu to the planar surface, where the uppermost Nitride-based etch stop layer can act as a polishing-stop indicator. CMP is a mechanically aggressive yet highly sensitive process, fine-tuned to achieve high yield on wafers of up to 300 mm in diameter [118].

Dielectric layer deposition is typically performed by variants of Chemical Vapour Deposition (CVD), with the metal barrier and seed layers deposited by sputtering, a form of Physical Vapour Deposition (PVD). Due to the significant difference in

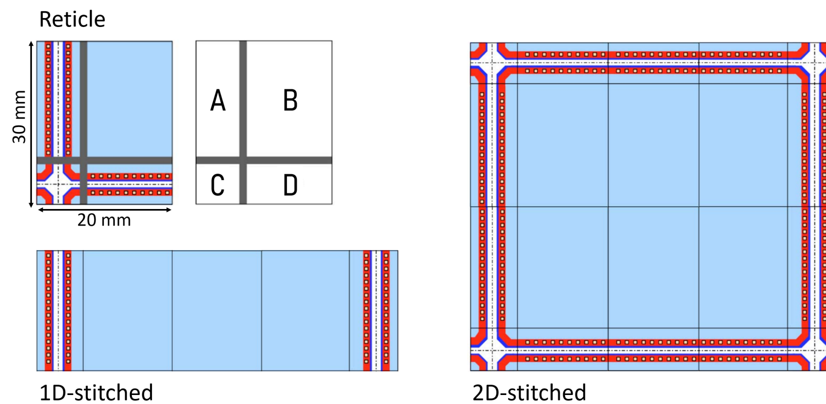


Figure 3.8: Illustration of the stitching concept to manufacture silicon devices larger than the design reticle size. Adapted from [125].

the Coefficient of Thermal Expansion (CTE) between copper and the dielectric (by roughly a factor of 10), substantial tensile stress is induced in the constrained copper during processing at temperatures above $\gtrsim 300°$C. Precise process control and e.g. device annealing are therefore required to prevent both void and hillock (growth of copper structures at grain boundaries) formation, otherwise detrimental to the device functionality (see also Chapter 7) [115, 119–123].

3.4.3 Stitching

The maximum design reticle size for the 65 nm technology node used in this work, and more generally for the ITS3 sensors, is about 20×30 mm^2 [124]. However, the largest monolithic sensor planes required for the ITS3, fabricated as a single continuous piece of silicon, measure 98×266 mm^2. A processing technique – stitching – is therefore employed for the first time in high-energy physics[5], to fabricate these large-area, 'wafer-scale' devices. The concept is illustrated in Figure 3.8.

In stitching, the photolithographic mask set is split into sub-frames (A, B, C, D), which are selectively exposed onto adjacent positions during the lithography process steps. This requires highly precise wafer stepping to align the exposures. At the abuttment boundaries, the metal wiring is designed such that interconnections are

[5]Stitching has been previously demonstrated for CMOS imaging sensors, e.g. in [125–128], and recently for a dedicated machine-learning processor termed Wafer-Scale Engine WSE [129].

joined, forming a connected circuitry over the entire stitched device area larger than the reticle frame size.

Designing such a device requires a periodic design in the central repeated area, and specific design rules for the metal lines to extend across stitching boundaries [130, 131]. As illustrated, stitched devices can be manufactured both as 1D stitched and 2D stitched chips. One crucial parameter is the functional device yield, as all individual stitched units need to be functional (at least to a certain degree, such as powering, see further below). To assess the feasibility of using stitching, test chips have been developed and are characterised in this work. The chip is introduced in detail in the next Section 3.5. The layout for the final sensor is discussed in Section 3.6.

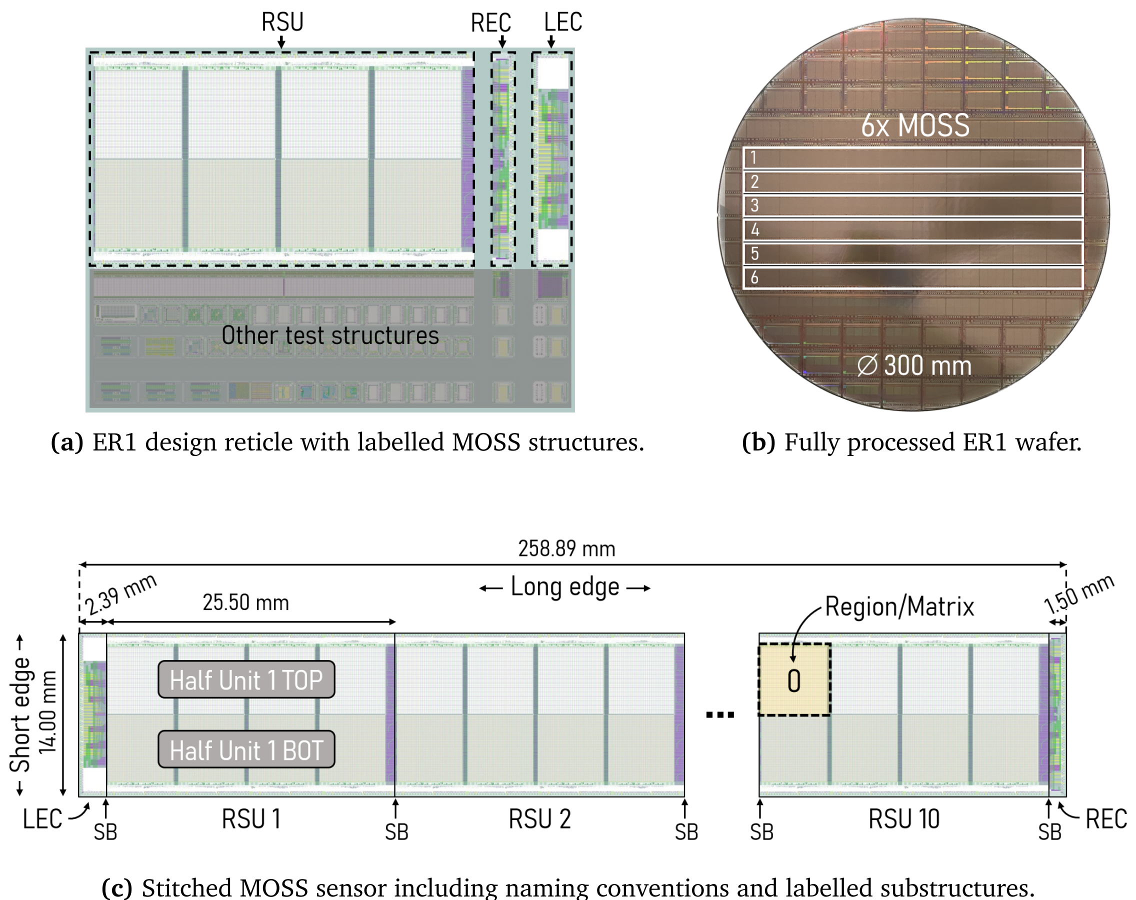
3.5 Monolithic Stitched Sensor MOSS

The MOnolithic Stitched Sensor (MOSS) is a prototype chip, manufactured in 65 nm TPSCo CMOS imaging technology. It is designed to study the yield and performance characteristics of a wafer-scale stitched sensor. This section discusses the design characteristics of the MOSS chip, as used in this work.

3.5.1 MOSS design and architecture

Stitching is employed to manufacture a monolithic pixel sensor of 14×259 mm^2 size. Three design structures of the Engineering Run 1 (ER1) design reticle (see Figure 3.9a) are used to construct one MOSS chip: the Left-Endcap (LEC), the Right-Endcap (REC), and the Repeated Sensor Unit (RSU). Ten RSUs are stitched together, and the chip is completed with each one LEC and REC on the left and right ends, respectively.

The MOSS sensor is a 1D stitched device. Each ER1 wafer contains 6 MOSS chips (see Figure 3.9b). A schematic of the MOSS sensor is shown in Figure 3.9c. Each RSU is split into a top and a bottom half, referred to as Half Unit (HU), with 4 pixel matrices (within 'regions') each. The top 4 pixel matrices have a pixel pitch of 22.5 μm and 256×256 pixels each, while the bottom 4 pixel matrices have a pixel pitch of 18 μm and 320×320 pixels. The analogue front-end design power densities of the top and bottom matrices are 7 mW cm^{-2} and 11 mW cm^{-2}, respectively.



(a) ER1 design reticle with labelled MOSS structures.

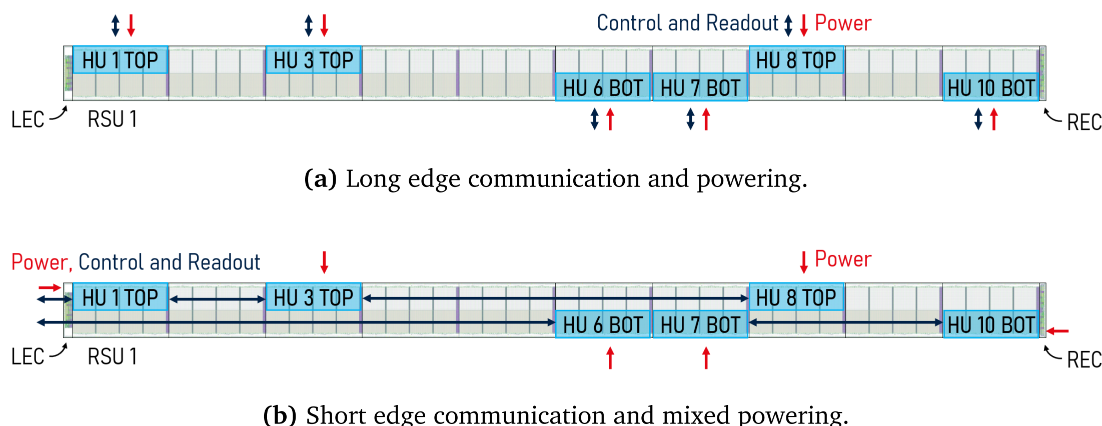
(b) Fully processed ER1 wafer.

(c) Stitched MOSS sensor including naming conventions and labelled substructures.

Figure 3.9: (a) ER1 design reticle with MOSS RSU, MOSS REC, MOSS LEC highlighted. (b) Fully processed ER1 wafer (300 mm diameter) with 6 MOSS chips in the centre. (c) Stitched MOSS sensor: ten RSUs are stitched together, completed with each one LEC and REC structure. Stitching Boundaries (SB) are indicated. Each RSU has a top and bottom Half Unit (HU) with each four pixel matrices. Wire-bond pads are located around all four edges of the chip for individual characterisation of HUs. The long and short edges of the chip are referred to as 'long edge' and 'short edge'.

In total, one MOSS chip has approximately 6.72 million pixels. Each of the 20 HUs (10 top HUs, 10 bottom HUs) can be operated independently by interfacing the top or bottom periphery wire-bond pads along the 'long edges' of the chip. This allows individual characterisation of HUs, and in case of failures, avoids rendering the entire MOSS unusable. Stitching across the RSU boundaries, exclusively on the metal lines of the sensor, allows the transfer of signals to and from the individual RSUs to the LEC (including power lines to the LEC and REC).

Two main operational schemes exist as schematically shown in Figure 3.10: (a) individual powering, control, and readout of the 20 HUs from each set of top/bottom



(a) Long edge communication and powering.

(b) Short edge communication and mixed powering.

Figure 3.10: (a) Powering, control, and readout of individual HUs via top and bottom wire bond pads along the long edge of the chip. (b) Control and readout of all, or individual HUs via the LEC (short edge). Powering via the long edge for all HUs. Optionally, RSU1 (HU1 TOP, HU1 BOT) and RSU10 (HU10 TOP, HU10 BOT) can be powered from the LEC and REC, respectively.

wire bond pads ('long-edge'), and (b) control and readout of the 20 HUs via the LEC, and powering via LEC and REC. Given the known limitations in the on-chip metal stack of the MOSS sensor in terms of voltage (IR) drops, additional powering via top and bottom wire bond pads is required for RSUs 2–9.

Unless otherwise noted, scheme (a), where each HU is interfaced with individually, is used throughout this work.

The functional block diagram of one HU is shown in Figure 3.11. Each pixel matrix has a dedicated analogue biasing block for the pixel front-end, as well as row and column steering for pixel control. Each matrix also features unique Digital-to-Analog Converter (DAC) control, pixel configuration, and region readout blocks, forming a 'region' that interfaces with the top-level slow-control and readout systems. Each matrix biasing block consists of 4 voltage DACs (vDAC, 8-bit) and 4 current DACs (iDAC, 8-bit). An on-chip multiplexer and monitoring output pads along the long edges of the chip allow for probing the output of each DAC. Each HU has $4(\text{regions}) \cdot [4(\text{vDAC}) + 4(\text{iDAC})] = 32$ DACs, resulting in 640 DACs per full MOSS chip. Override pads allow for externally biasing the chip front-end via wire-bond pads in case of DAC failures.

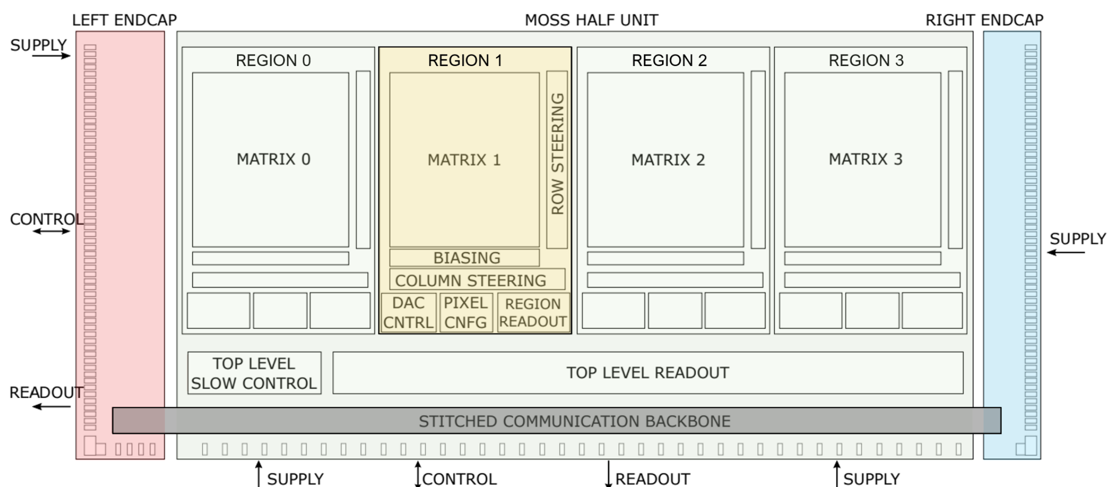


Figure 3.11: Functional block diagram of one bottom HU of the MOSS sensor, including LEC and REC. Adapted from [6].

The pixel front-end is derived from the Digital Pixel Test Structure (DPTS) prototype [107] and shown for reference in Figure 3.12. Each pixel has its own front-end, which is biased from one biasing block for each matrix (every pixel front-end is biased the same way within one matrix), and a binary discriminator (a detailed discussion is found in [106]). The configurable threshold level is therefore common to all pixels of the same matrix. Analogue pulsing circuitry is included for characterisation and calibration. The charge injected is varied by setting VPULSEH, and is at maximum $Q = 258\ \text{aF} \cdot 1.2\ \text{V} \simeq 1932\ e^-$ (design value). Digital in-pixel logic controls digital and analogue pulsing, and pixel masking (see Appendix A.2).

The binary hit information is stored as the coincidence of the pixel discriminator output and a global 'strobe' pulse (of programmable duration). Each region readout zero suppresses the matrix hit information and relays the row and column addresses to the peripheral top readout. Hit data is transferred off-chip by the top-level readout – either via the long-edge wire-bond pads (per HU) or via the stitched backbone from the LEC. Per-HU configuration is done via a synchronous 1-bit-in, 1-bit-out serial slow-control interface, with 40-bit commands specifying command type, ID, address, and value. Readout occurs via an 8-bit synchronous parallel port upon request. The word size is therefore 8 bits, and an event consists of a header, ID,

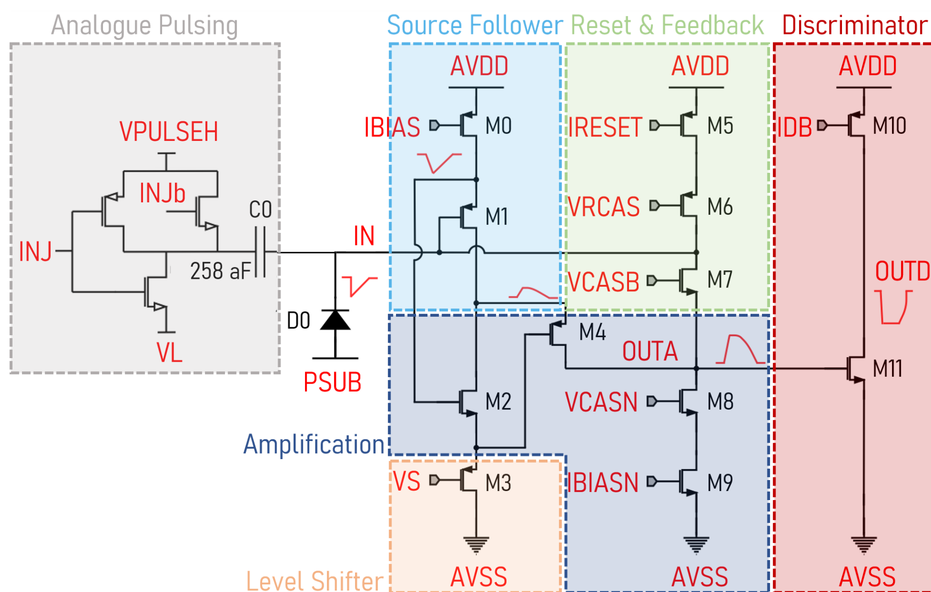


Figure 3.12: Annotated analogue pixel front-end. Biasing voltages and currents are generated by on-chip DACs, and are common to all pixels in one matrix. Adapted from [107].

data, and trailer [22]. The chip is operated at a clock speed of 33.3 MHz, translating to a 33.3 MB/s data rate for data readout.

Each HU has 8 power domains, described in Table 3.1. The analogue domain (AVDD, AVSS) supplies power to the analogue blocks, powering the pixel front-end. The digital domain (DVDD, DVSS) powers the readout and control circuitry. The IOVDD net supplies the peripheral level shifting circuitry (the off-chip signalling is at 1.8 V potential compared to the 1.2 V on-chip signals). The backbone domain (BBVDD, BBVSS), consisting of one pair of supply lines for each of the top and bottom halves of the MOSS chip, powers the stitched communication backbone that enables signalling between the HUs and the LEC. The chip substrate (PSUB) can be reverse-biased (typically between 0 and –1.2 V) to increase the electrical field in the detection volume of the chip. The backbone power nets and chip substrate span the full chip, across stitching boundaries. In contrast, all other power nets are designed to remain independent between HUs when each HU is powered via its dedicated long-edge wire-bond pads.

Additional power distribution lines for AVDD, AVSS, DVDD, and DVSS for each HU and spanning the full chip exist and are routed to the LEC and REC. The on-chip

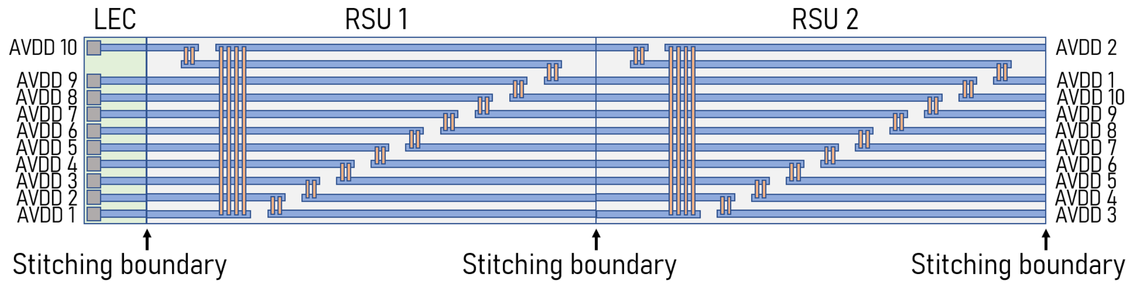


Figure 3.13: Concept of line-hopping to distribute power and signals across stitched structures from the LEC to individual RSUs/HUs. As an example, the AVDD power distribution for the first two RSUs (out of 10, and only one upper half of the chip) is illustrated: The power lines effectively 'hop' by one power line, and each unit receives the same-numbered AVDD net at the uppermost power line (as illustrated for AVDD 2). The hopping itself is periodic, and each RSU has the exact same design (such that the same reticle structure can be used for manufacturing the stitched sensor). Horizontal and vertical metal lines are routed on two different layers of the chip metal stack, and can therefore cross. For the MOSS chip, this powering scheme is not sufficient, given the known *IR* loss in the metal stack (only RSU1 from LEC and RSU10 from REC can be powered in this way, and all other HUs must be powered from the long edge). Signalling over the backbone to the LEC, however, functions the same way and is fully functional.

metal stack is, however, only designed to sufficiently power RSU 1 from LEC and RSU 10 from REC, according to the powering scheme illustrated in Figure 3.10b. The concept of addressing or powering a specific HU from the LEC or REC is achieved by 'line-hopping', as illustrated in Figure 3.13 for one arbitrary supply line (AVDD) for 10 HUs (2 of which are illustrated). In this concept, each HU shifts the lines by one using two metal layers, enabling the interfacing of individual HUs from a single end [22]. The signalling and readout from the LEC is implemented in the same way (including signal regeneration on-chip where needed).

Test-out pads for each HU allow reading out probe signals via a programmable

Table 3.1: MOSS power domains and nominal supply voltage.

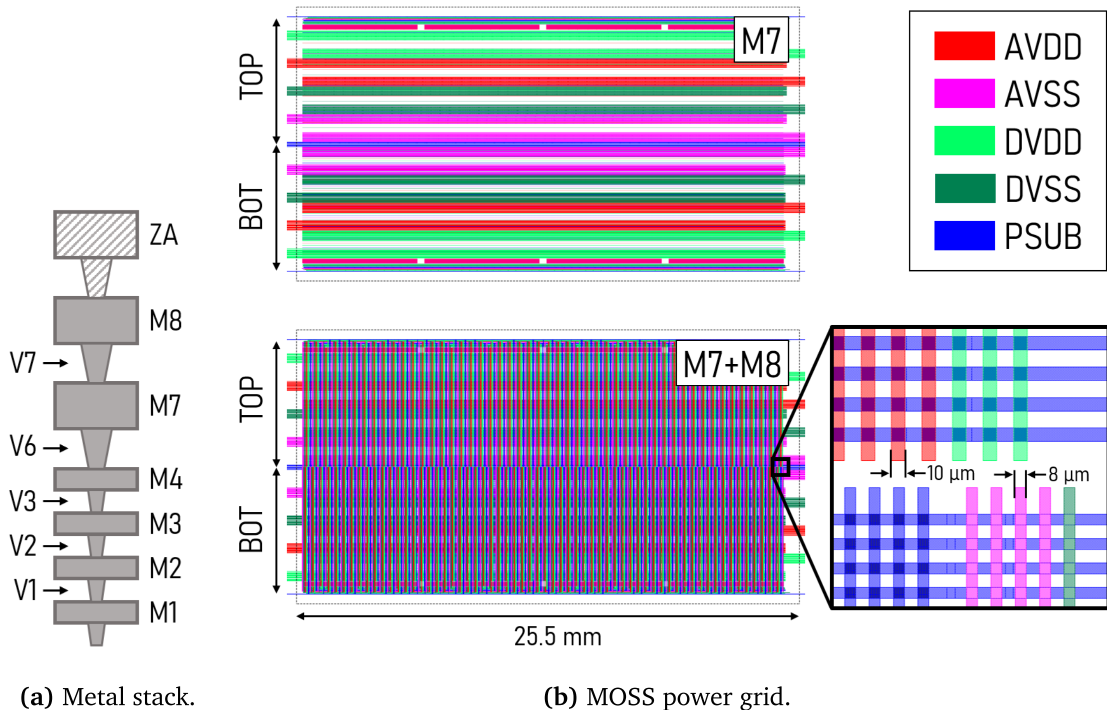
Ground net	Supply net	Functional domain	Nominal voltage [V]
AVSS	AVDD	Analog	1.2
DVSS	DVDD	Digital	1.2
DVSS	IOVDD	Digital I/O	1.8
BBVSS	BBVDD	Backbone	1.2
BBVSS	BBIOVDD	LEC digital I/O	1.8
all	PSUB	Chip substrate	0 or -1.2

multiplexer. A total of 2192 wire bonds electrically interconnect the MOSS chip to the readout PCB (see Section 4.3). One MOSS sensor has a total of 390 digital I/Os, 480 analogue I/Os, and 107 individual supply nets. The highly granular design of the chip allows for in-depth characterisation of chip performance and yield parameters (independently for each HU), but it requires sophisticated and complex test systems.

Overall, 24 ER1 wafers were produced (6 MOSS sensors each). During transport, thinning, and dicing, 4 wafers were destroyed, leaving 20 wafers for characterisation.

3.5.2 Metal stack and power grid

The copper metal stack of the MOSS chip is manufactured in a dual-damascene process as described in Section 3.4, and schematically shown as a cross-section in Figure 3.14a. There are 6 copper layers, M1–M4, M7, M8, and one aluminium-based grid on top of the chip (ZA layer). The ZA grid is on PSUB potential (except at the wire-bond pads). Wire-bond pads for electrical interconnection of the chip are fabricated on the ZA layer. Layers M7 and M8 are approximately twice as thick as M1–M4. Their dielectric composition also differs, although most intermetal dielectrics are lower-k ($2.5 \lesssim \epsilon_r \lesssim 3.9$) variants of SiO_2. Exact layer dimensions and dielectric composition are proprietary information and cannot be disclosed. The power grid, supplying each MOSS HU uniformly with power, is routed on the uppermost two copper layers of the copper metal stack: layers M7 and M8. A top-down view extracted from the chip design is shown in Figure 3.14b, where power lines are routed horizontally on layer M7 and vertically on layer M8. Vias (V7) connect layers M7 and M8 at the intersections of power nets within the same domain. All other crossings are insulated by the dielectric layer. The line-width and spacing ratio of M7/M8 in the bottom HU compared to the top HU is 4/5 (i.e., 0.8). In the detailed view shown in Figure 3.14b, which focuses on the boundary between the top and bottom HUs, the metal line widths are 10 µm and 8 µm, respectively. While the routing density varies between the top and bottom halves, the overall metal density is the same. Very conservative line spacing, exceeding the minimum design rules, was chosen to maximise yield.

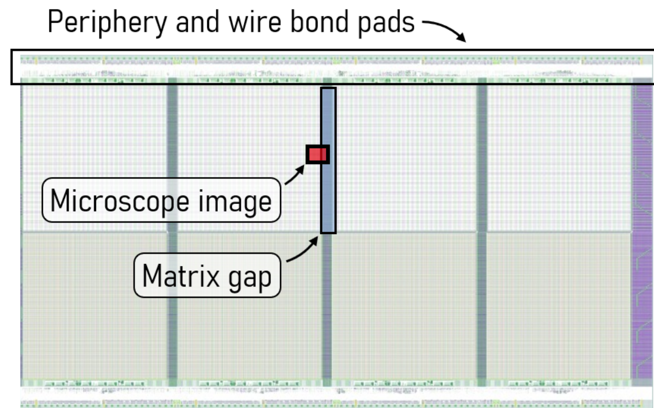


(a) Metal stack.

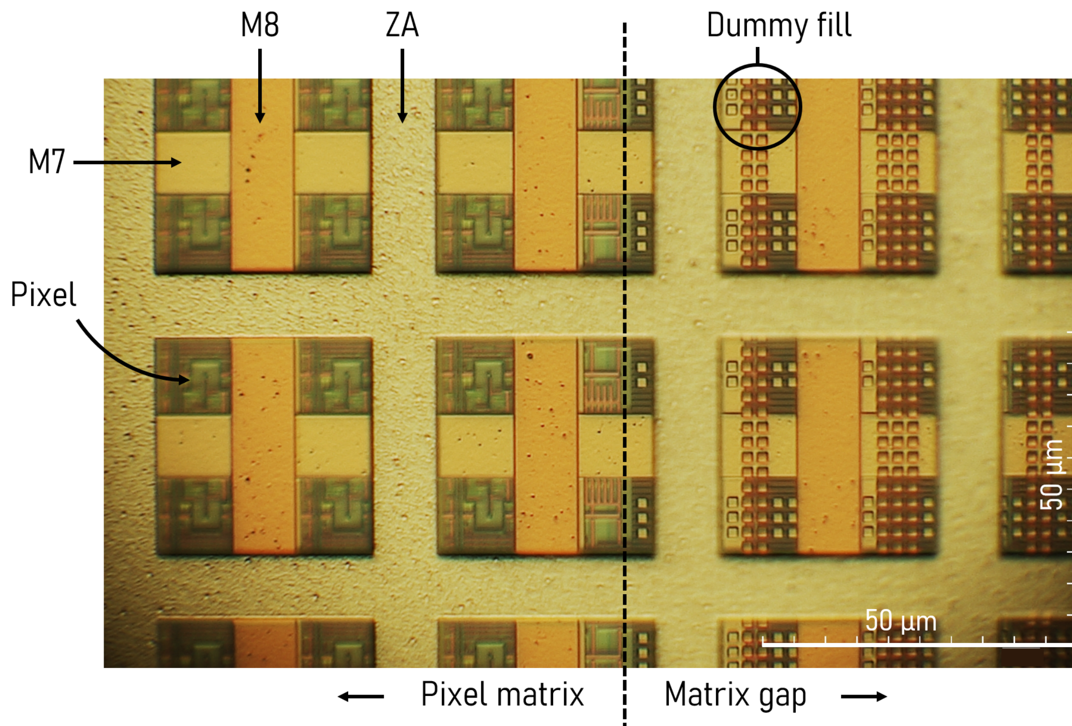
(b) MOSS power grid.

Figure 3.14: (a) Schematic view of metal stack layers of the MOSS chip in a cross-section view. (b) MOSS power grid of one RSU: The power nets are routed as a grid on metal layers M7 (horizontal) and M8 (vertical). The detailed view on the bottom right shows the higher routing density and smaller linewidth of the bottom HUs, compared to the top HUs of the MOSS.

A top-down microscope image of a fully processed MOSS sensor is shown in Figure 3.15. Given the transparency of the dielectric, it is possible to see metal layers down to M1. To the left side of the dashed line, part of a pixel matrix is shown. Individual pixels can be discerned, where no metals are covering the collection diode to minimise the material budget locally. The aluminium-based ZA grid, as the uppermost layer, is clearly visible, with the power grid routed on the M7 and M8 copper layers indicated. The ZA grid, M7, and M8 layers cover large parts of the M1–M4 metal routing. On the right side of the dashed line, the matrix boundary, i.e., the area between two pixel matrices, is partially shown. In this area, only M7 and M8 are routed. To keep the metal density across the MOSS sensor uniform, dummy metal fill is added, which has no electrical connection to any power or signal-carrying metal lines. The M7 and M8 metal lines visible have a linewidth of 10 µm and spacing of 12 µm (cf. Figure 3.14b).



(a) Illustration of microscope image location, and matrix gap locations on one RSU.



(b) Microscope image of pixel matrix and the matrix gap region of a MOSS top HU.

Figure 3.15: (a) Pixel matrix gap areas, periphery, and microscope image location illustrated on a MOSS RSU. (b) Top-down microscope image of a MOSS sensor. The boundary of the pixel matrix (left) and the pixel matrix gap (right) is shown. The ZA grid on top (Al), vertically routed M8 (Cu) and horizontally routed M7 (Cu) are clearly visible. Lower metals and single pixels can be distinguished. In the gap region (in between two pixel matrices/regions), a small patterned 'dummy fill' maintains a uniform metal density for a given metal layer across the sensor.

3.6 ITS3 sensor layout

The final sensor is conceptually similar in design to the MOSS; however with finer powering granularity, increased fill factor, and a revised readout architecture. This sensor is introduced here, as simulations in Chapter 8 are based on the ITS3 geometry using this layout. At the time of writing, the sensor is in its final review stages before submission to the foundry[6]. It remains a 1D-stitched chip incorporating LEC, RSU, and REC design blocks, as shown in Figure 3.16. One fully independent structure, consisting of one LEC, 12 RSUs, and one REC, is termed a 'Segment' as illustrated in Figure 3.17. 1D stitching is chosen, allowing for a variable number of segments to be cut out from a wafer ('diced'), depending on the sensor layer of the ITS3 being fabricated. For the outermost layer L2, the full available area comprising 5 segments is diced to form one sensor plane. Two such layers are required, one for the top L2 half-layer and one for the bottom L2 half-layer. For sensors used in layers L1 and L0, four and three segments are diced from the wafer, respectively. In total, 6 wafers are required to produce the sensors for a complete ITS3 barrel. For L1 and L0 sensors, since not all segments are needed, cuts can be optimised by selecting among adjacent segments with the highest yield (two and three permutations, respectively; see also Figure 8.14b).

Each RSU contains 12 tiles, which can be switched on, biased, and read out independently. This provides a much higher design granularity, with the intention of switching off malfunctioning tiles in the event of a fault. To support this, a very conservative design with high yield is chosen for the switching circuitry, which in turn allows for more aggressive routing and optimisation at the pixel level.

One segment contains 12 RSUs × 12 tiles = 144 tiles, each with 444 × 156 pixels per tile (20.8 × 22.8 μm pixel size), and a total of approximately 10 million pixels per segment. The effect of malfunctioning tiles, which have to be turned off, on the ITS3 detector performance and geometry optimisation is further discussed in Chapter 8.

[6]For Engineering Run 2 (ER2). The final detector layers will be manufactured in an additional ER3 submission, which allows for minor bugfixes and removal of testing pads to maximise the fill factor, but without forseen changes to the geometrical and conceptual sensor layout.

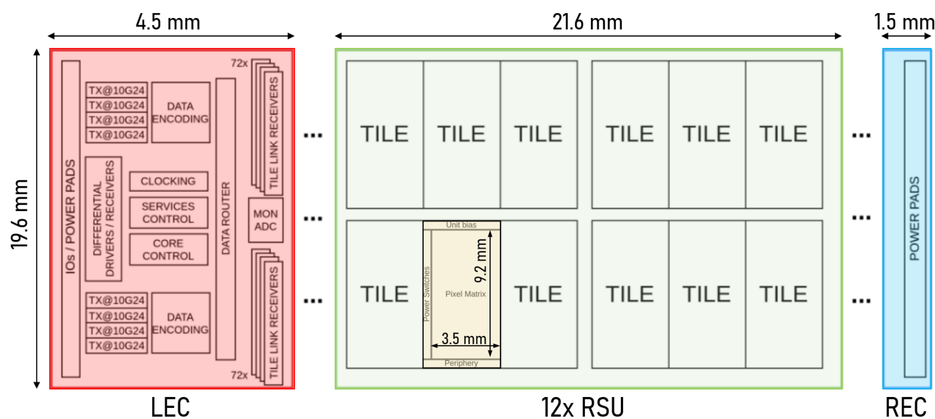


Figure 3.16: Design structures for the ITS3 sensor layout. Adapted from [6].

The fill factor of the active sensor area is not 100%, due to cutting areas between segments and areas surrounding each tile that are not covered with pixels. This has been accounted for in the simulations. A fill factor of $\sim$ 93% is currently achieved, with an estimated value of $\gtrsim$ 95.5% for the final iteration of the chip. A more detailed discussion of the current status of this sensor development can be found in [6, 132–134].

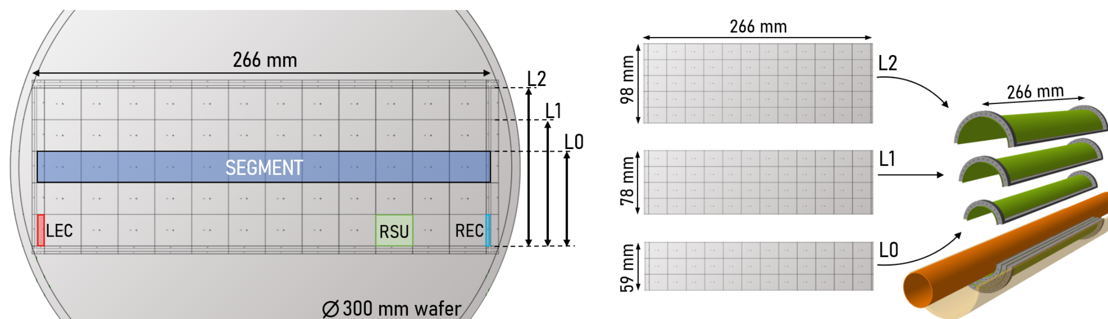


Figure 3.17: Wafer map of the ITS3 sensor layout. One segment consists of one LEC, 12 RSUs, and 1 REC. Three, four, or five adjacent segments are cut out to produce sensors for ITS3 layers L0, L1, and L2, respectively. Six wafers are needed for all sensors in one ITS3 barrel. Adapted from [135].

4

MOSS Handling and Mounting

The processed, 300 mm diameter ER1 wafers containing the MOSS sensors are received from the foundry with a thickness of approximately 700 μm. They are thinned to a thickness of 50 μm and diced (cutting of single MOSS sensors) at an external company [136]. These thinned and diced wafers are delivered on a wafer tape on a wafer frame. To handle the release of the 50 μm-thick structures, measuring 14×259 mm^2, as well as to pick them up and transfer them into a custom-designed storage box, dedicated tooling was developed. In addition, the carrier Printed Circuit Board (PCB) was specifically designed to allow through-PCB glueing, and specialised procedures and tools are employed to mount the MOSS sensor onto the carrier PCB. After mounting the MOSS sensor on the PCB, wire bonding is performed to connect the sensor to the PCB electrically. This PCB is used for the full electrical and functional characterisation of the MOSS sensor throughout this work. The custom tooling and procedures are described below.

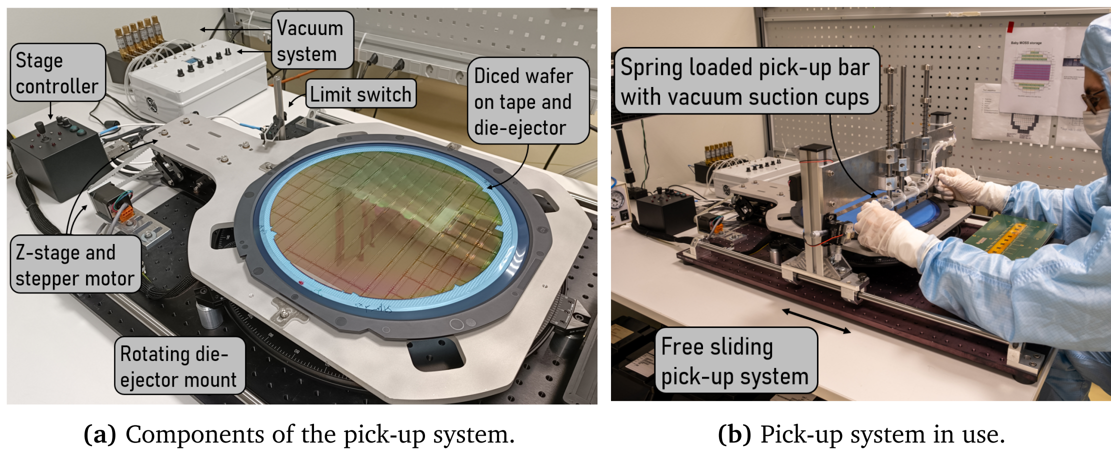
4.1 Pick-up system

The thinned and diced ER1 wafers are delivered mounted on a standard 300 mm wafer carrier. To release the single diced MOSS sensors (and other design structures) from the blue wafer tape (see Figure 4.1a), a commercial die ejector is used [137].

Custom tooling was developed to exploit this device. The die ejector consists of a circular aluminium disk with a patterned surface as shown in Figure 4.1c. The surface features a pyramidal pattern with a 1.27 mm (0.05 in) pitch and is perforated with evenly spaced holes approximately 1 mm in diameter to allow for vacuum application. After the wafer carrier is placed on top (see procedure below), the aluminium plate is heated to 70 °C with an attached heating element. This softens the glue between the wafer carrier tape and the diced silicon sensors. After approximately 15 min, a vacuum is applied through the holes in the patterned aluminium plate. The blue wafer carrier tape is then sucked into the trenches of the pyramidal patterned aluminium disk. This step releases the diced silicon chips from the blue tape, which are ready to be picked up.

Given the large dimensions and flexibility of the MOSS sensors (the sensor deforms under gravity, given the 50 µm thickness), it is not possible to use standard manual picking procedures, e.g. using a pick-up pen. The pick-up system is shown in Figure 4.1. It consists of multiple parts with different functionalities:

- Wafer frame carrier: An aluminium frame on a custom motorised z-stage receives the wafer frame carrying the thinned and diced wafer. This allows for slowly lowering the wafer frame onto the die-ejector while keeping both surfaces parallel.
- Die-ejector: Mounted on a rotational plate, the die ejector is placed in the middle of the pick-up system. The wafer frame is lowered onto the die-ejector, and the motorised rotational stage allows alignment of the MOSS sensor (and other structures) with the vacuum pick-up bar.
- Sliding pick-up stage: The pick-up stage is mounted on linear bearings, allowing movement both horizontally and vertically. The horizontal movement is free-wheeling and has a locking mechanism. The vertical movement is spring-loaded and returns automatically to the 'up' position when letting go. The vertical movement has an adjustable lower height limit, to avoid crashing the pick-up suction cups into the wafer. These movements are performed manually by the operator. Dedicated knobs to grab onto are located on the vacuum pick-up bar.



(a) Components of the pick-up system.

(b) Pick-up system in use.

(c) Vacuum pick-up bar with MOSS chip in mid-air. Patterned die-ejector visible through blue tape.

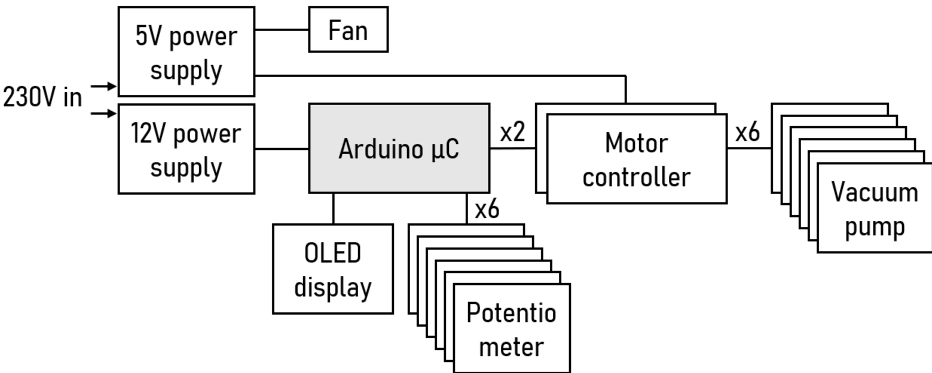
Figure 4.1: Picture of the pick-up system with annotated components (a), and in use (b). (c) Close-up view of the pick-up bar featuring five bellowed vacuum suction cups mounted on a spring-loaded and vertically travel-limited bar, used to safely pick up the chip.

- Vacuum pick-up bar: To pick up the MOSS sensor from the wafer tape (after being released from the blue wafer tape with the die-ejector), five bellowed vacuum suction cups [138] mounted on an aluminium bar are lowered and brought in contact with the chip. The vacuum suction then lifts the chip from the tape. Each line has an individual vacuum supply, such that loss of vacuum in one line does not cause an unintended release of the chip. The chip is now held in place securely and is moved horizontally towards the operator. The chip is then lowered onto a custom storage tray (located on a manually operated z-stage), and the vacuum and chip are released by pressing a button conveniently located on the vacuum bar (see red button in Figure 4.1c). A guide laser (mounted on a dual-axis goniometer) is used as a visual aid when lowering the vacuum

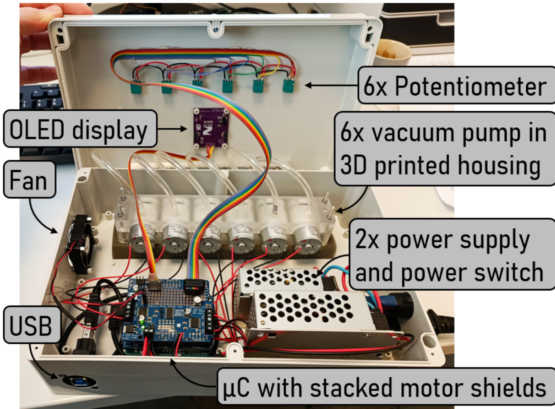
suction cups onto the chip surface, indicating where contact will be made.

- Vacuum system: The custom vacuum system is designed around 6 (5 used + 1 spare) miniature vacuum pumps driven by DC-motors [139] as shown in Figure 4.2b. The pumps are held in place by a 3D-printed structure and mounted inside the control box, decoupled by a foam layer to reduce vibration. Each pump's speed, and thus its flow, is regulated via a potentiometer and motor controllers, which are interfaced by an Arduino Leonardo [140]. Two motor controllers [141], each with four channels, are stacked (6 out of 8 total used). The operating state of each pump (as a percentage of maximum speed) is shown on a display. A small cooling fan is included to stabilise the vacuum pump temperatures inside the enclosure. The block diagram of the system is shown in Figure 4.2a. An external valve system, shown in Figure 4.2c, allows switching all vacuum channels at once via the button mounted on the pick-up bar. The valves used are normally open 3-way solenoid valves. The vacuum is always 'on' when lowering the pick-up bar onto the wafer for pick-up. After moving the chip above the storage plate or carrier PCB, the chip is released by switching all valves in parallel to the second position, connecting the vacuum lines to atmospheric pressure while cutting off the vacuum pumps. A power MOSFET is used to switch the valves simultaneously, and high-speed diodes protect the MOSFET and power supply from voltage spikes induced by the valve coils.

- Z-axis and rotation controller: A joystick-operated control system was developed using an Arduino-compatible microcontroller board [142], as illustrated in Figure 4.3. Vertical movement of the joystick controls the lowering and raising of the wafer frame onto the die-ejector, by actuating a z-stage driven by a stepper motor [143]. The motor is controlled via a driver board [144], which is interfaced with the microcontroller. The movement speed is regulated through a potentiometer, whose output is read by the Arduino's ADC. Two safety switches limit the z-stage travel distance and are wired to the interrupt pins of the Arduino, stopping movement immediately if triggered. For rotational movement,



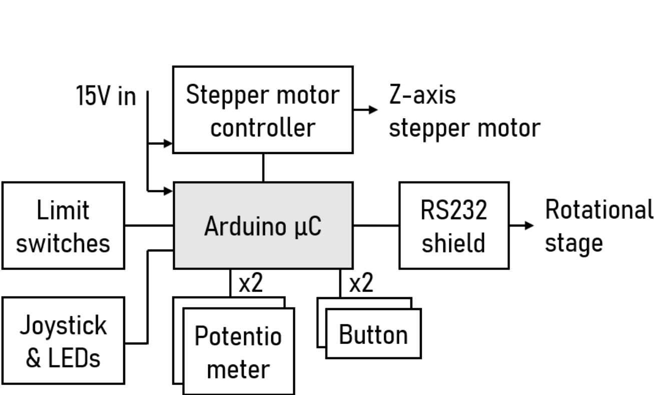
(a) Functional block diagram.



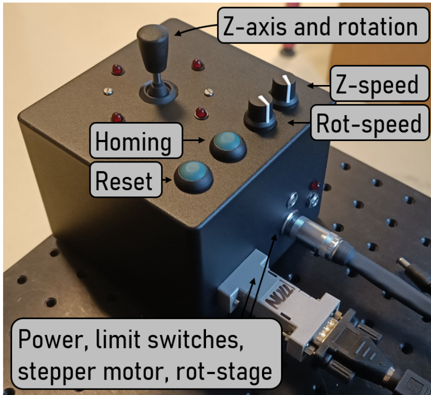
(b) Opened vacuum pump system.

(c) Valve block.

Figure 4.2: (a) Block diagram of the vacuum pump system illustrating the used components. (b) Opened vacuum pump system enclosure and annotated components. Each miniature vacuum pump is controlled and set independently from 0 to 100% flow rate. (c) The external valve block has an integrated power supply and power MOSFET for switching the 6 solenoid 3-way valves in parallel, releasing the picked-up chip with the push of a button.



(a) Functional block diagram.



(b) Finished controller.

Figure 4.3: (a) Block diagram of the z-axis and rotational movement controller illustrating the used components. (b) Finished controller ready for integration in the pick-up test system. An external 15 V wall power supply with a DC-barrel connector is used.

a motorised rotational stage [145] is used, which is interfaced via an RS232 interface and a corresponding shield mounted on the Arduino. The horizontal movement of the joystick translates to left and right rotation of the die-ejector and wafer carrier frame (allowing for fine-tuning the alignment of the MOSS and other sensors to the vacuum pick-up bar). The rotational speed is again controlled by reading a potentiometer input. Two additional buttons allow for resetting and homing the stages, respectively.

The MOSS storage trays are custom-designed, each accommodating up to 6 MOSS sensors from a single wafer. The pick-up system is also used to pick up the MOSS sensors from the storage trays and place them onto the carrier PCB for mounting. Two pad-wafers (dummy wafers with only top layer metal processing) were used to test, validate, and commission the pick-up system. Overall, 84 MOSS chips (and additional structures) from 14 wafers were picked successfully, without any breakage.

4.2 Mounting

Mounting the MOSS sensor to the carrier PCB is non-trivial. The dimensions and wire bonding pads along all four edges of the chip make conventional glueing techniques unfeasible. A new procedure was therefore developed, as described below.

4.2.1 MOSS carrier PCB

The MOSS carrier PCB has the function of securely holding the chip in place, and providing electrical interconnections for testing distributed to five high-density connectors (see also Section 4.3). The PCB is purely passive, with decoupling capacitors, series resistors and eight NTC probes as only components (excluding an I^2C addressable unique ID chip for automatic board identification). For a reliable wire-bonding process, the chip (especially the chip edge, where the bond pads are located) needs to be mechanically supported. The PCB has four important features:

- Gold-plated central chip support area: The Electroless Nickel Immersion Gold (ENIG) surface finish must be of the highest quality, and without any nodules,

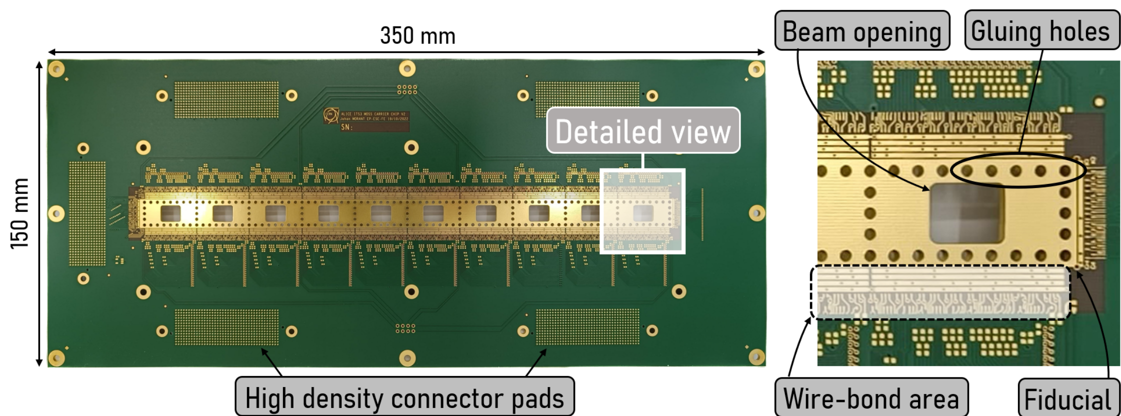


Figure 4.4: The bare MOSS carrier PCB without components. The chip is mounted on the gold-plated surface in the centre. A detailed view is shown on the right. The beam opening in the centre (one per RSU) is clearly visible, with glueing holes located around the opening. Wire bond pads are located around the perimeter of the chip mounting area. One fiducial, used to align the chip during mounting, is indicated.

which can introduce mechanical stress when the chip is placed on top. The drilling and milling direction (from top to bottom) is specified for fabrication, avoiding sharp overhangs. Through-PCB vias were ultimately filled to avoid solder seepage onto wire-bonding pads during component mounting.

- Beam opening: Each of the 10 RSUs of the MOSS sensor has a centred rectangular beam opening below in the PCB. This allows for test beam measurements without multiple scattering effects introduced by the PCB, relevant for high precision spatial resolution and detection efficiency determination.

- Glueing holes: Around the chip support area and each RSU boundary, 193 2 mm diameter through-PCB glueing holes are placed. These glueing holes are filled with glue during mounting and hold the chip in place, mechanically connecting the backside of the chip with the hole walls in the PCB.

- Fiducial holes: Four small holes with 0.5 mm diameter are placed at the corner positions of the chip, and allow for precise alignment under the microscope during the mounting procedure.

The PCB is industrially manufactured, and one out of three vendors delivered the required surface quality for mounting the MOSS chips. The electronic components,

including connectors, are mounted at the CERN electronics workshop. After cleaning and visual inspection, the MOSS carrier boards are ready to receive a MOSS chip.

4.2.2 Chip mounting

The steps for mounting the MOSS chip on the carrier PCB are outlined below. After chip placement and alignment as illustrated in Figure 4.5, the chip is glued and cured (see Figure 4.6).

1. The carrier PCB, fully assembled with components and connectors, is mounted on an aluminium jig with screws (stand-offs).

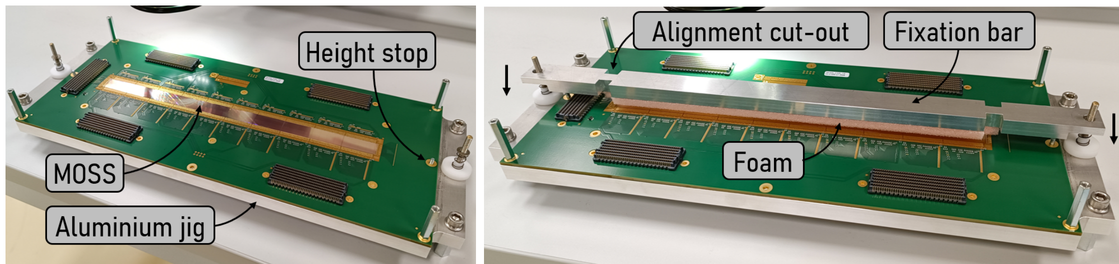
2. The chip is picked from the storage box and placed in the centre of the carrier PCB using the pick-up system.

3. The jig with the carrier PCB and MOSS chip on top is placed under a microscope, and the chip corners are coarsely aligned with the fiducial holes (see Figure 4.5a). Clean room swabs are used to gently move the chip.

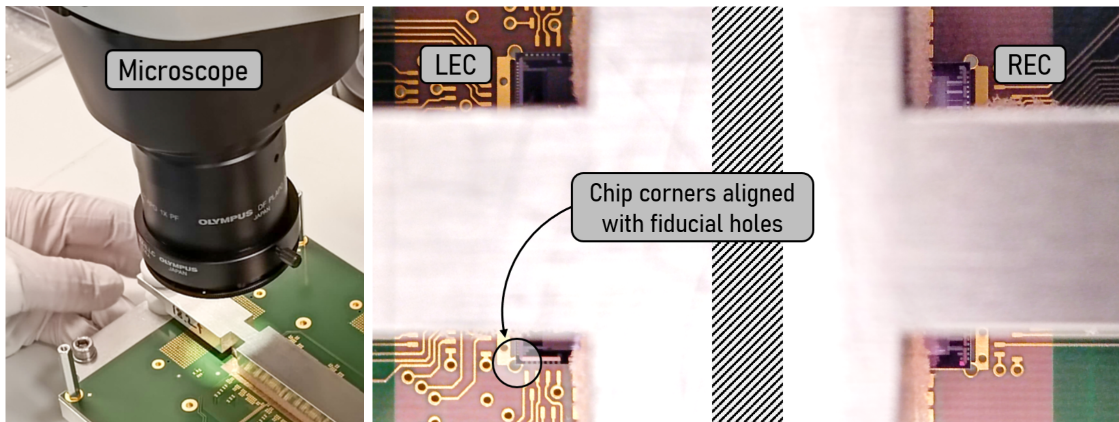
4. A spring-loaded aluminium bar ('fixation bar') with Electro Static Discharge (ESD) safe foam on the bottom is placed above the chip. The foam is not in contact with the chip yet. Custom nuts on either side of the fixation bar allow to slowly lower the bar and bring the foam into contact with the chip. The bar is designed to permit lateral adjustment in both axes over a range of approximately ± 2 mm, allowing for fine adjustment of the MOSS position on the carrier PCB (see Figure 4.5b).

5. The chip corners are then precisely aligned with the fiducials on the PCB using the optical microscope. Cut-outs in the bar allow looking at the chip corners from above in a top-down view (see Figure 4.5c).

6. Once alignment is accurate, the nuts are tightened until the bar hits a pre-defined mechanical height-stop. This ensures uniform and repeatable pressure on the chip, securely holding it in place while limiting the pressure. A chip



(a) MOSS on the carrier PCB and jig. **(b)** Fixation bar above MOSS sensor, before tightening.



(c) Microscope alignment. **(d)** MOSS chip corners aligned with fiducial holes.

Figure 4.5: Alignment of the MOSS chip on the carrier PCB: (a) Placement of the MOSS chip on the fully assembled PCB (the five high-density electrical connectors are visible in black). The PCB is mounted on an aluminium jig. The MOSS sensor corners are coarsely aligned with the fiducial marks. (b) The fixation bar is placed above the chip. The bar is spring loaded and tightened down (indicated by arrows), until the foam is touching the sensor. (c) Cut-outs in the fixation bar allow for seeing the chip corners from above, allowing fine adjustment of the chip position by gently moving the bar under a microscope. (d) Fully aligned MOSS sensor. The fixation bar is tightened down until it hits a pre-set height stop.

alignment accuracy of $O(100\ \mu\text{m})$ along both axes is achieved. After tightening, the alignment is checked (see Figure 4.5d).

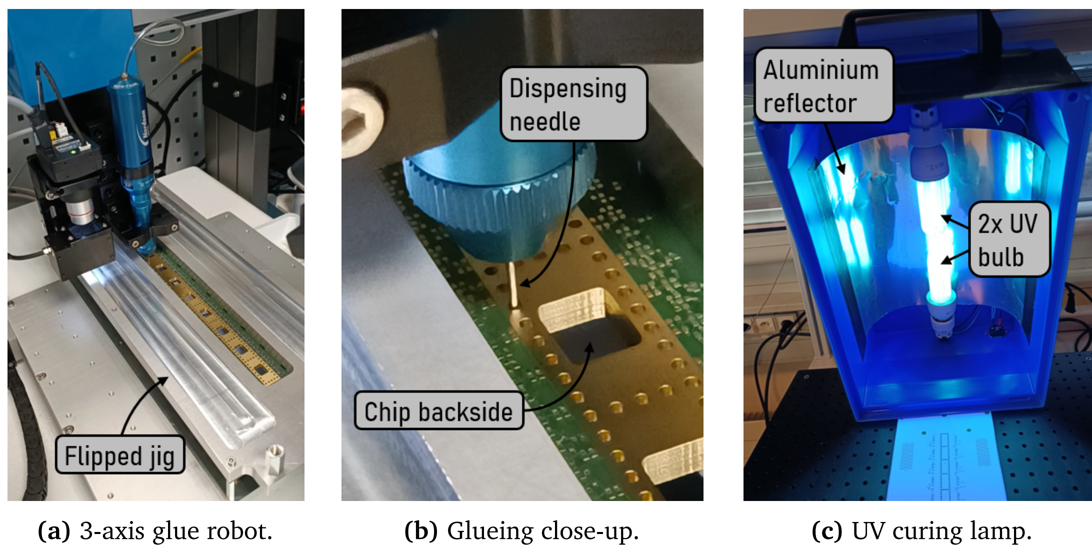
7. The entire aluminium jig, holding carrier PCB, MOSS sensor, and tightened-down fixation bar is then flipped by 180 degrees, such that the bottom of the PCB is pointing upwards, and transferred onto a 3-axis glue robot (see Figure 4.6a). The rectangular cutout in the centre of the jig allows access to all 193 glueing holes. The sturdy aluminium jig ensures uniform height of the carrier PCB and MOSS chip.

8. Using a custom-developed program and procedure, the glue robot is set up. As the dispensing needle height changes with glue syringe replacement, the height is manually set by lowering the needle onto the PCB with a clean room wipe in between. Upon contact, the height is set. Horizontal alignment is performed using pre-defined fiducials and a camera mounted on the glue robot head. Using the fiducials, a correction to horizontal misalignment (small rotation) is automatically applied, ensuring the correctly centred needle position in each glueing hole. The needle is lowered into each hole to a distance of 0.5 mm from the chip's backside. A predefined amount of UV glue (by setting the needle diameter of 0.5 mm, the dispensing time of 0.15 s, and a pressure of 1 bar) is then automatically dispensed in every glueing hole (see Figure 4.6b). UV glue [146] is used due to its re-workability and its ability to retain a degree of flexibility after curing. This is desirable, as it allows for a slight potential deformation of the chip without leading to stress or possibly cracking in the large silicon sensor. Glueing studies are detailed further below.

9. After all glueing holes are filled, the jig is unmounted from the glue robot and placed under a custom UV lamp. The range of maximum absorption of 350–380 nm of the UV glue required for curing matches the emission spectrum of commercially available UV bulbs with a typical wavelength of 368 nm, as commonly used for insect lamps. A box with two UV fluorescent bulbs and a metallic reflector was built to illuminate the full glueing area of the MOSS carrier PCB at once (see Figure 4.6c). The UV glue is cured for 30 min, mechanically connecting the backside of the chip to the carrier PCB.

10. Wire bonding is performed in two steps (limited by the working area of the wire-bond machine, given the size of the chip, see Section 4.3).

Three sets of custom mounting jigs were manufactured to enable parallel processing and speed up assembly: chip placement and alignment, chip glueing, and chip curing can be performed simultaneously for up to three MOSS sensors. Overall, 83 out of 84 MOSS sensors were successfully mounted using this procedure. In



(a) 3-axis glue robot. **(b)** Glueing close-up. **(c)** UV curing lamp.

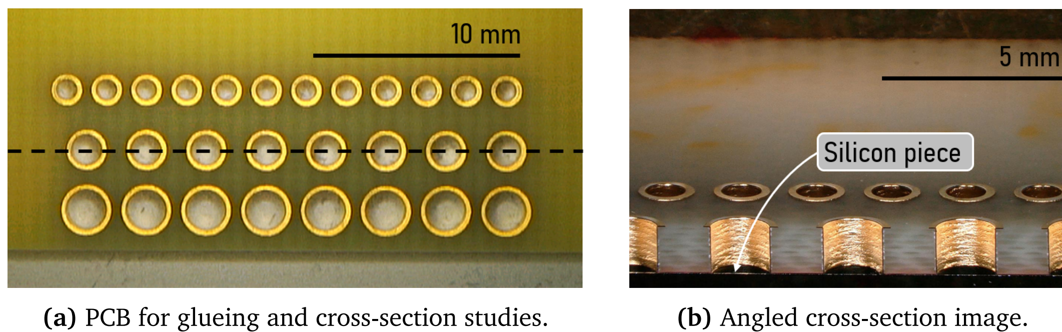
Figure 4.6: (a) A MOSS sensor mounted in the aluminium jig, flipped such that the backside of the MOSS carrier PCB and MOSS sensor is facing upward. The glueing holes are exposed. (b) Detail of the automated glueing process with the glueing needle placed inside one glueing hole. The chip backside is visible through the beam opening in the carrier PCB. High needle placement accuracy is required. (c) The UV curing lamp uses two UV bulbs and an aluminium reflector.

one case, the height setting step of the glueing needle was omitted by distraction, and the glueing needle damaged the sensor (9 out of 10 RSUs of this sensor could still be characterised).

Gluing validation

Glueing tests were performed to ensure a consistent and reliable connection of the chip to the PCB. UV glue was chosen over epoxy glue:

- The UV glue does not harden until exposed to direct UV light. This allows for easy adjustments and re-workability – for example, to salvage a sensor. This did not occur during the MOSS assembly.
- No glue mixing is required for UV glue, and the UV glue viscosity stays constant (during the 8-month shelf life). This saves time and ensures a consistent volume of glue is deposited by the glue dispensing tool over time.
- The UV glue chosen exhibits flexibility when cured. This is beneficial in case the PCB and therefore chip slightly bends, or expands and contracts during



(a) PCB for glueing and cross-section studies.

(b) Angled cross-section image.

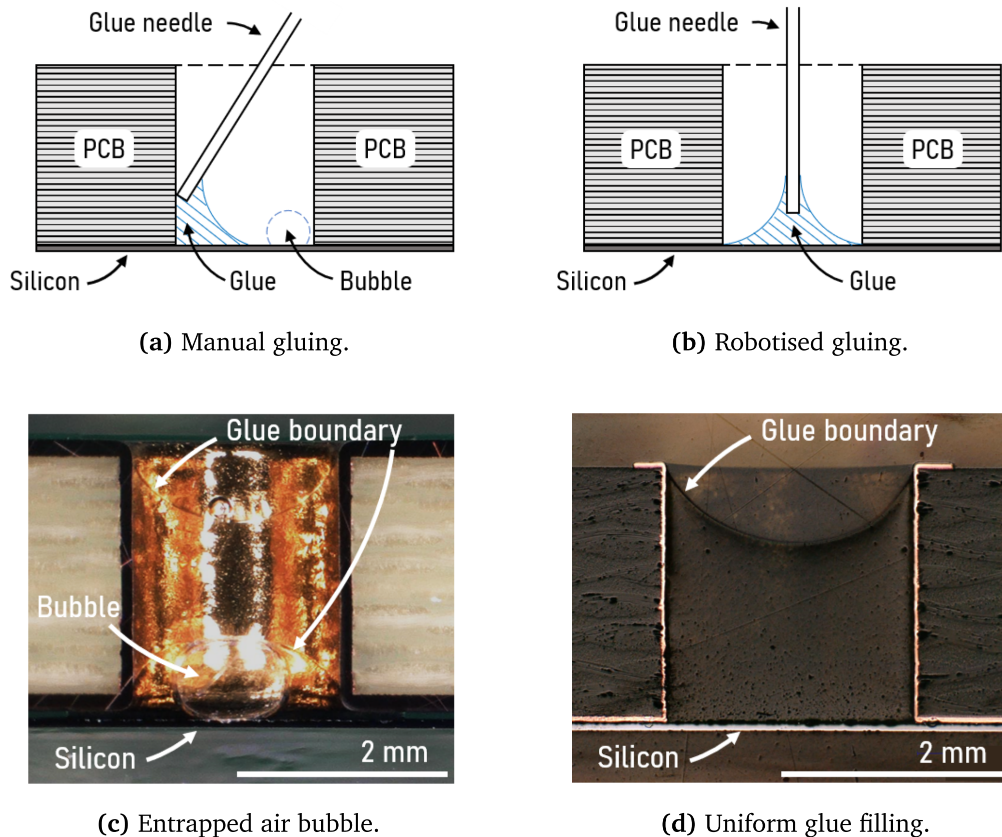
Figure 4.7: (a) PCB for glueing studies with three rows of glueing holes with diameters ranging from 1 mm to 2 mm. The cross-section line is indicated. (b) Angled cross-section image at the location indicated in (a). The glueing holes are filled with transparent, cured UV glue. The silicon test piece is visible at the bottom.

handling or operation, decreasing the stress in the chip itself.

To ensure the reliability of the process, test samples were manufactured, and cross-sections were fabricated. The testing PCB with multiple glueing holes and hole diameters is shown in Figure 4.7a. A silicon test piece of about 1×2 cm^2 is used. After glueing and curing, the entire sample is encapsulated in epoxy resin. Mechanical abrasion and polishing down to the glueing hole enabled inspection of the glueing quality (see Figure 4.7b). Initially, glueing was performed manually, with a pressure-driven syringe, and the glue needle placed into the glueing hole by hand under a microscope as illustrated in Figure 4.8a. This procedure consistently resulted in air bubbles trapped at the chip surface due to the capillary effect within the cylindrical glueing hole (see Figure 4.8c). Air bubbles potentially lead to issues during wire-bonding if the adhesion of the chip to the PCB is insufficient (as the chip is slightly pulled on), or if the MOSS chip were to be placed under vacuum (e.g. for single event effect tests) the pressure difference could lead to a silicon break-through given the 50 µm thickness of the chip.

Therefore, and to speed up assembly while drastically reducing the risk of damage to the chip by contact with the needle, I proposed the use of a 3-axis glue robot. The use of the glue robot enabled precise control of glue deposition, resulting in repeatable results and uniformly filled glueing holes. The needle is vertically inserted in the centre of each glueing hole, at 0.5 mm distance from the chip surface, and glue spreads

out uniformly, avoiding trapping of air bubbles (see Figure 4.8b and Figure 4.8d). Out of 20 analysed glueing holes in two samples, 100% show no entrapped air bubbles when using the automated procedure, confirming its repeatability.



(a) Manual gluing.

(b) Robotised gluing.

(c) Entrapped air bubble.

(d) Uniform glue filling.

Figure 4.8: Illustration of the manual glueing process where the glueing hole wall is used as a guide (a), and resulting air bubble entrapment given the capillary effect on the UV glue visible in the cross-section image (c). The 3-axis glue robot allows for centred, repeatable placement of the glueing needle (b), with uniform glue deposition without air bubble entrapment, visible also in the corresponding cross-section image (d).

4.3 Interconnection

After glueing and curing, the MOSS sensor is wire-bonded in two steps (1140 + 1052 wire-bonds), as the working area of the bonding machine limits the bonding process to one half at a time. Wire-bonds are located around all four edges of the MOSS chip. A detailed view of a wire-bonded MOSS chip from the LEC (rotated 180 degrees) is shown in Figure 4.9. Wire-bond pull tests were performed on a pad-wafer chip (non-

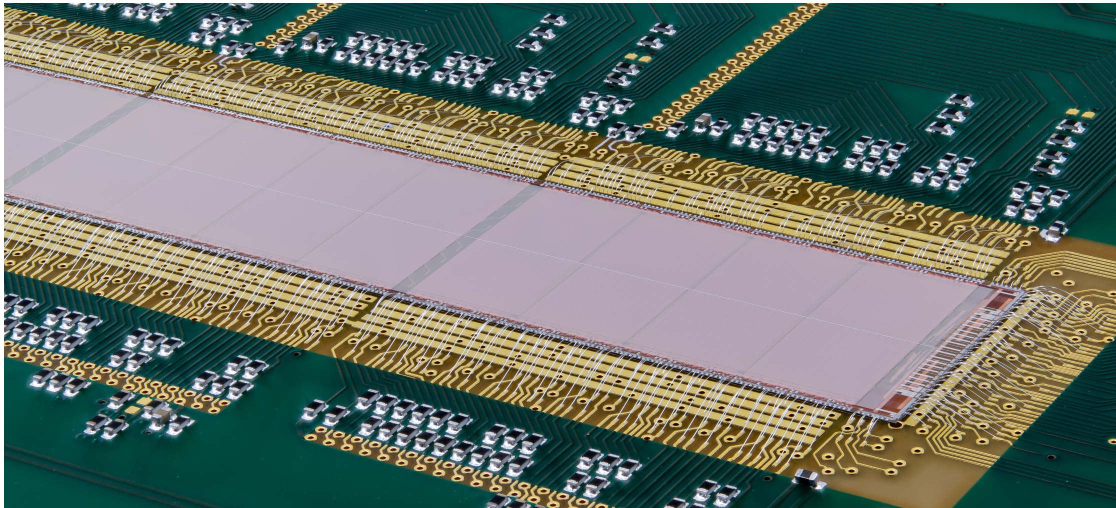
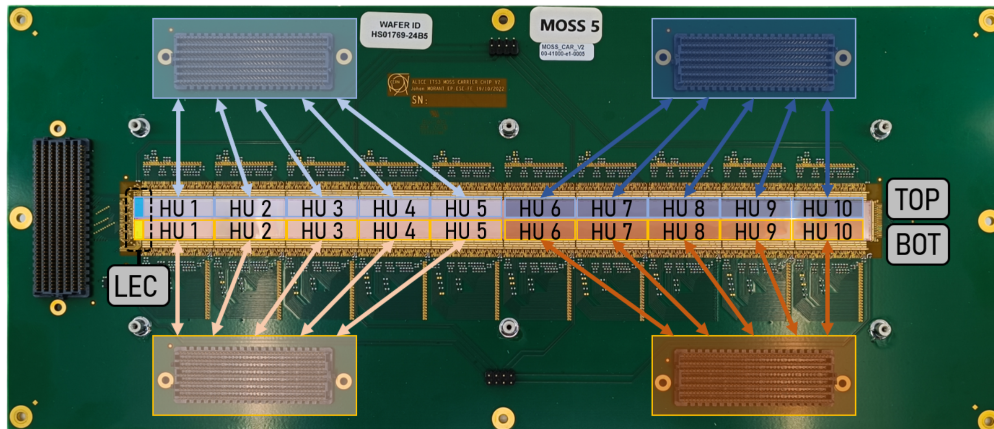


Figure 4.9: Close-up, focus-stacked image of a wire-bonded MOSS sensor. Wire-bonds are located along all four chip edges. The LEC, RSU1, RSU2, and part of RSU3 are visible (the board is rotated 180 degrees compared to the usual orientation).

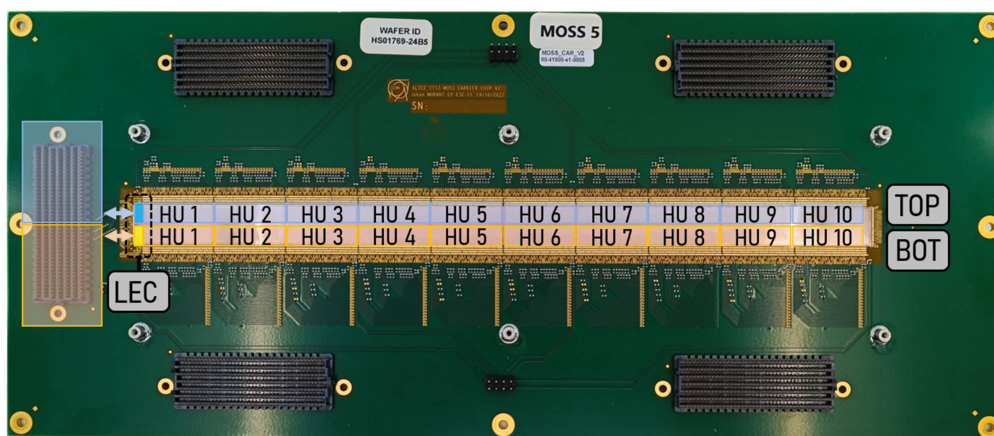
functional silicon dummy with uppermost metal layer for mechanical tests). Overall, 30 pulls were made, with all tests passing, breaking at a mean of 10.71±0.22 g.

The MOSS sensor is now ready for electrical and functional characterisation using the measurements described in Chapter 5. Five high-density connectors with 560 pins each are located around the MOSS carrier PCB, used to power, control, readout, and monitor the MOSS chip. As discussed in Section 3.5.1, two powering and communication schemes exist. As illustrated in Figure 4.10a, four of the high-density connectors allow access to each HU individually via the long edge of the MOSS chip. To test the LEC communication, the fifth high-density connector is used as illustrated in Figure 4.10b, allowing the interface of all top and bottom HUs. In this communication scheme, power still needs to be provided to individual HUs (of RSU2–RSU9) via the long edge connectors. The global PSUB and backbone power can be supplied via any connector[1].

[1]The separate backbone nets for the top and bottom half of the chip are supplied from any top and bottom half connector, respectively



(a) Individual HU communication from the long edge connectors.



(b) HU communication via the short edge connector and LEC.

Figure 4.10: Fully mounted and wire-bonded MOSS sensor on the MOSS carrier PCB. The two interfacing schemes are illustrated: (a) Communication via the long edge of each HU, and (b) communication via the short edge (LEC).

5

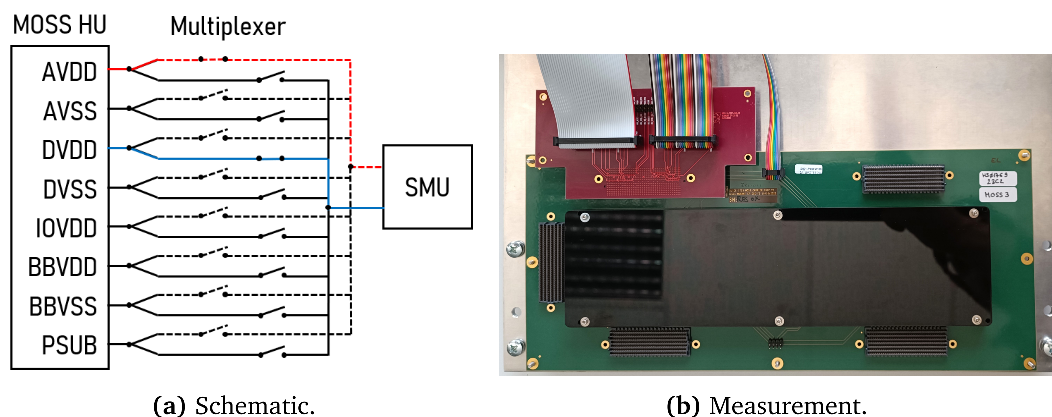
MOSS Test Systems and Measurements

To characterise the MOSS sensor performance (excluding test beam measurements), three main test systems were developed. These will be discussed below.

5.1 Impedance measurement

Impedance measurements are performed to detect potential short circuits between power nets, possibly indicating manufacturing defects. Each HU (Half Unit) of the MOSS chip is treated as an individual sensor with independent power domains. Eight power nets per HU are driven from the corresponding bonding pads on the top/bottom edge of the chip (including the global backbone and PSUB nets spanning the entire sensor, see Section 3.5.1).

To measure the impedance between any of the power net combinations, a Source Measure Unit (SMU) [147] in combination with a channel multiplexer [148] is used. The circuit diagram is shown in Figure 5.1a for one HU. For a given net combination, the corresponding multiplexer channels are closed programmatically and connected to the SMU output (highlighted example in red and blue for the power net combination AVDD–DVDD). In the testing setup, one breakout board allows access to 5 HUs at a time (see Figure 5.1b). The multiplexer is therefore configured for 5 HUs in parallel to minimise testing time. Altogether 5 (HUs) · 8 (Power Nets) ·

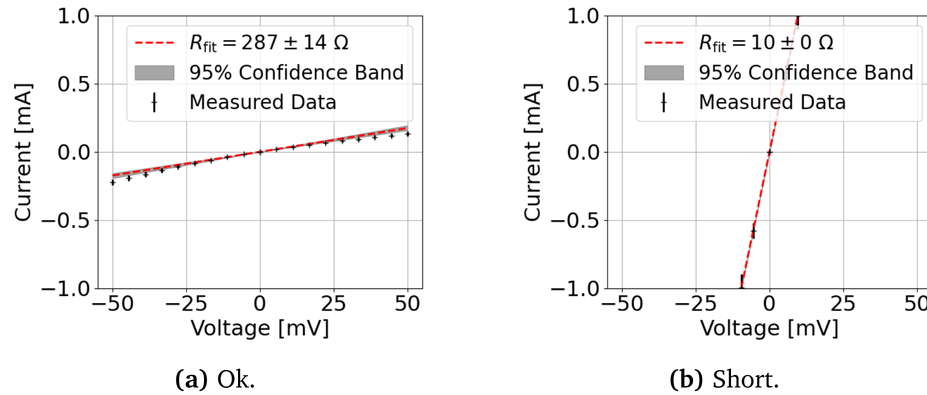


(a) Schematic. **(b)** Measurement.

Figure 5.1: (a) Schematic diagram of the impedance measurement setup for one HU. An example configuration for the AVDD–DVDD path is highlighted in red and blue. (b) MOSS testing PCB with light cover, and the breakout board for the impedance measurement (red) connected on the top left. 5 HUs are measured subsequently with the setup, before the breakout board is plugged into the next high-density connector. The flatband cable in the centre of the board is used to measure the board temperature via NTCs.

2 (Connections) = 80 (Multiplexer Channels) are required[1]. Two flatband cables are used to connect the breakout board to the channel multiplexer. An additional flatband cable (connector located between the high-density connectors) provides access to two NTC thermistors on the PCB, enabling the recording of the board's temperature during the impedance measurement. The multiplexer instrument features an onboard Analogue-to-Digital Converter (ADC), allowing the direct measurement of the (converted) NTC temperature on two separate channels. Diode structures between chip power nets lead to non-ohmic temperature-dependent measurements. Although a temperature correction was not required for the type of analysis conducted, the measured temperature enables error estimation given fluctuating lab temperatures on the order of 10% (see below). The measurement is performed with a protective light cover over the chip. Illumination of the chip leads to an offset current for some power net pair combinations, generated by diode structures acting as parasitic photodiodes. The lead wire resistance contributes on the order of 1.0–1.5 Ω to the value measured, depending on multiplexer routing.

[1]In the present implementation, the number is reduced to 60, given the global PSUB net, and global top and bottom backbone nets not requiring per-HU interfaces.



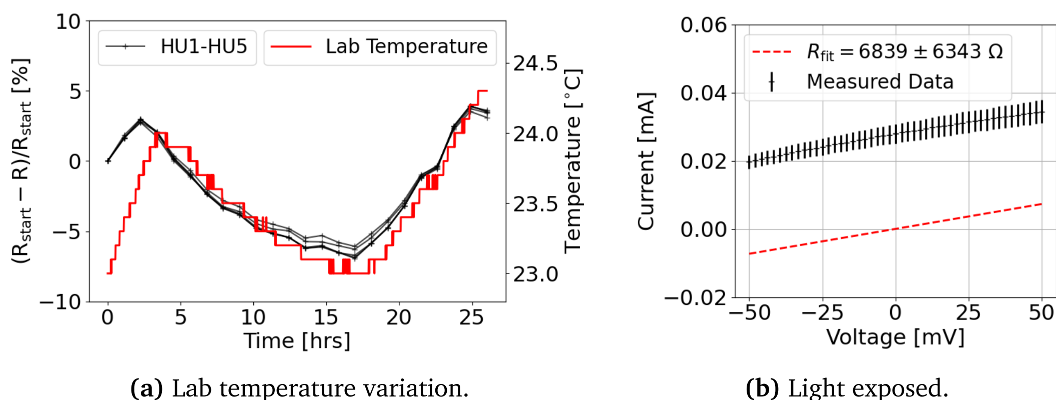
(a) Ok.

(b) Short.

Figure 5.2: Example of impedance measurements between power net pairs, and classified as (a) 'ok', and (b) 'short'.

The test system is controlled with custom Python code that interfaces the instruments (SMU, channel multiplexer). Parameters, including voltage range, current limit, step size, and automatic data visualisation, are configured at run-time. During the series testing of the MOSS chips, the values quoted below were fixed.

For each HU, 8 power net pairs are tested, which equates to $\binom{8}{2} = 28$ possible combinations per HU, and a total of 560 measurements per MOSS chip. For each power net pair, a small voltage is applied and ramped in steps of 5 mV between 0 to –50 mV and 0 to +50 mV. This range was chosen to stay below a threshold at which transistor structures become highly conductive. The current is measured for each voltage step using the SMU. If a current of 1 mA is exceeded, the measurement is stopped. The resulting current-voltage diagram gives an indication of the impedance, and a linear fit is performed to approximate the resistance between the nets under test using Ohm's law $U = R \cdot I$. Based on the distribution of impedance measurements, a global cut of 30 Ω was chosen, where a lower value is classified as a 'short' (see Section 6.2.1). Examples of power net pairs considered 'ok' and 'shorted' are shown in Figure 5.2. Note: The term impedance is chosen over resistance, given the non-ohmic nature of silicon devices. However, an ohmic approximation is considered reasonable within the small voltage range applied in this test. Non-shorted net-pairs may exhibit rather low $O(\gtrsim 50\ \Omega)$ impedances given the cumulative effect of on-chip diode connections, leakage paths, and interacting device circuitry.



(a) Lab temperature variation.

(b) Light exposed.

Figure 5.3: (a) Variation of measured impedances for one power net pair combination (AVSS-AVDD) on 5 HUs with varying lab temperature over a span of 24 hours. (b) Effect of exposing the sensor to light during the impedance measurement: On a subset of power net pairs a current is generated, here shown for the AVSS-PSUB power net pair on one HU, and generated by (parasitic) diode structures effectively acting as photo diodes.

For reference, the effects of varying lab temperature and measuring a light exposed sensor are shown in Figure 5.3a and Figure 5.3b, respectively: The impedance on one power net pair combination (AVSS–AVDD) on 5 HUs, repeatedly measured over a span of 24 hours is shown together with the lab temperature. A clear correlation is observed with an impedance variation of $O(10\%)$ for a temperature change of $O(1.5°\mathrm{C})$ for the given power net pair. This temperature dependence, which varies across power net pairs, does not affect the detection efficiency of shorts, and is therefore not discussed further. Similarly, the effect of exposing the sensor to light is not relevant, as tests were performed with a light protective cover. However, the measurement of a light-exposed sensor reveals a shift of the I-V curve. Specifically, a non-zero current is observed even when no external voltage is applied, due to parasitic photodiode effects. As a result, the curve is offset by this zero-point current, and the resulting linear fit – constrained to pass through the origin – fails, as illustrated in Figure 5.3b. Neither of these effects impacts further measurements.

5.1.1 Wafer probing

On a subset of 12 wafers, impedance measurements were (additionally) performed using a probe card. The probe card gives access to all pads of one HU at a time. There are no active components on the probe card, similar to the testing PCB. Impedance

measurements were performed before the wafers were thinned and diced. One HU is contacted at a time, before moving the wafer to the next HU position. The impedance measurement follows the same procedure as described above, using the same type of measurement instruments. An image of the wafer prober with installed probe card and loaded ER1 wafer is shown in Figure 5.4. The probe card allows to fully operate one MOSS HU, and is not limited to impedance or powering measurements.

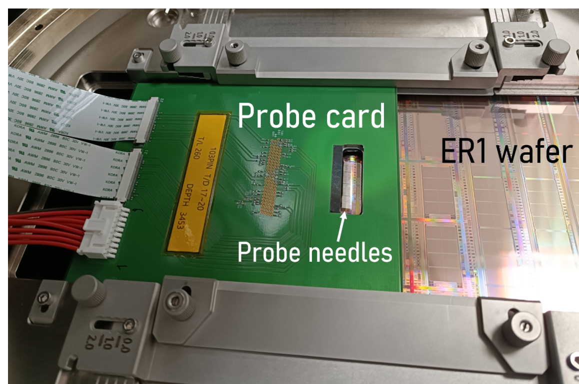
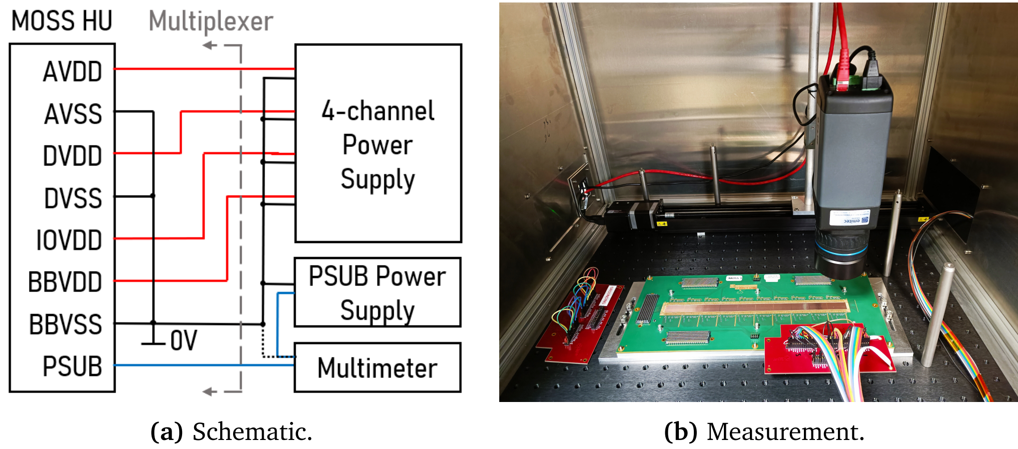


Figure 5.4: ER1 wafer mounted in the wafer prober with the installed probe card. One HU can be contacted at a time. All pads are contacted and routed off the probe card via flat band cables.

5.2 Power ramping with thermal camera analysis

The initial power-on of the MOSS chip is performed with a dedicated setup. For each HU, each of the power nets is ramped up to nominal voltage sequentially. This approach was chosen to better understand potential failure modes on a limited number of sensors, as discussed below. The ramp-up sequence was empirically defined while considering chip design constraints. All ground nets (AVSS, DVSS, BBVSS) are connected together (off-chip). First data were taken with the substrate (PSUB) on the same ground potential (0 V). At a later stage, the power ramp-up included a PSUB ramp to –1.2 V as a first step. All ramps are performed in 12 steps, starting at 0 V (0.1 V step size for 1.2 V nominal voltage). Current limits are set programmatically for each power net, and the power ramp is stopped if limits are exceeded. Current limits were iteratively increased after gaining understanding of the chip behaviour, as discussed in Section 6.3. The full sequence is given in Table 5.1.



(a) Schematic.

(b) Measurement.

Figure 5.5: (a) Schematic diagram of the power ramping setup for one HU. If PSUB is at 0V, the PSUB power supply is not used, and PSUB is connected to ground via the multimeter (see black dashed line). A multiplexer is used to measure 5 HUs connected via a breakout board. (b) Power ramping setup. The red breakout board is used to provide power to the MOSS HUs. A thermal camera on a linear stage is positioned over the HU under test. A light-shielded enclosure protects the chip from being illuminated.

Step I is omitted when PSUB = 0 V. In Step II, the DVDD net is ramped to its nominal voltage and subsequently set back to 0 V. In Step III, the analogue domain (AVDD) is powered up, and kept powered on if the nominal voltage is reached within current limits. Step IV then ramps up the digital domain (DVDD), such that both AVDD and DVDD are powered on. Similarly, IOVDD and BBVDD are powered up in Step V and Step VI, respectively. The HU under test is fully powered after completing Step VI within current limits. In Step VI the HU is kept powered on for 5 seconds to test for a non-fluctuating current consumption. An example of a ramp-up sequence is shown in Figure 5.6 at PSUB = 0 V. In this example, a transient high current is observed during the ramp-up of the analogue domain (AVDD). After an ohmic (linear) current increase starting at 0 V, a sharp drop is visible, and the current follows the expected power-on

Table 5.1: MOSS power ramp-up sequence.

Power net	Step I	II	III	IV	V	VI	VII
PSUB [V]	0 – -1.2	0/-1.2	0/-1.2	0/-1.2	0/-1.2	0/-1.2	0/-1.2
DVDD [V]	0	0 – 1.2	0	0 – 1.2	1.2	1.2	1.2
AVDD [V]	0	0	0 – 1.2	1.2	1.2	1.2	1.2
IOVDD [V]	0	0	0	0	0 – 1.8	1.8	1.8
BBVDD [V]	0	0	0	0	0	0 – 1.2	1.2

shape. Such a non-repeatable current drop following an ohmic increase is termed a 'burn-through'. It indicates a low-impedance power net pair caused by a short fault in the chip metal stack as discussed later in Chapters 6 and 7. It should be noted that the power consumption and final steady-state currents vary significantly between HUs and MOSS sensors at this stage. This variability arises because proper chip reset and configuration can only be achieved by writing to multiple configuration registers.

The schematic connection diagram is shown for one HU in Figure 5.5a. Currents are measured with the 4-channel power supply [149] for the power nets AVDD, DVDD, IOVDD, and BBVDD (return currents are not measured). The PSUB current is measured with a digital multimeter [150]. If PSUB is applied, an additional power supply is used. A multiplexer is used to switch between connections of 5 HUs, each contacted with the breakout board plugged to one high-density connector of the MOSS testing PCB. The setup and instruments are controlled by custom Python code, and parameters for power ramp-up are defined in a configuration file and loaded at runtime. Five HUs are measured automatically, before an operator moves the breakout board to the next position on the MOSS testing PCB.

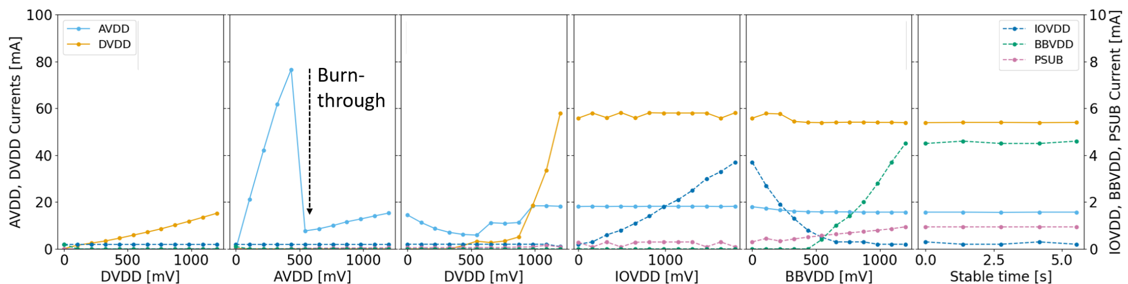


Figure 5.6: Power-on ramp of one HU at PSUB: 0 V. Each plot corresponds to a step as outlined in Table 5.1. The rightmost plot shows the stable currents with all nets at nominal voltage. A high transient current, or 'burn-through', is visible during the AVDD ramp-up. This corresponds to the 'burn-through' of a short circuit structure in the MOSS chip metal stack.

5.2.1 Hotspot localisation with a thermal camera

Following the observation of low impedance power net pairs in the impedance measurement, and non-repeatable high transient currents during power ramp-up, a thermal camera [151] was added to the power ramping setup (Figure 5.5b). The

thermal camera allows for the identification of areas of high current consumption through localised heat signatures. Mounted on a motorised linear stage, the position of the thermal camera is software-controlled. With a field-of-view corresponding to about one RSU, the thermal camera is automatically positioned above the HU under test (inside the light-shielded enclosure). The thermal camera sensor has 640 × 480 pixels, resulting in an approximate 50 μm resolution using a close-up infrared lens [152].

Short-circuit failures are located as hotspots visible during the power ramp-up of a given HU. Correlation with the impedance measurement and power-on curves enables conclusions to be drawn about potential failure modes.

A semi-automated procedure was developed to identify the hotspots, determine their position, and map the locations to the MOSS sensor design coordinates. It was observed that hotspots often disappear, relating to a burn-through of a local short, permanently changing the impedance of the affected net from low to high (cf. Figure 5.6, and discussion in Chapter 6).

The following steps are performed to extract the coordinates of a hotspot location matched to the MOSS global coordinate system. Thermal camera images are recorded with an interval of 150 ms during the power ramp. These are greyscale (one channel) images which are stored with a 14-bit depth. The duration of a full power ramp is approximately 70 s, with approximately 530 images stored. An automated procedure is used to detect images with hotspots and extract the coordinates semi-automatically. A user confirms the suggested locations, which are visually marked on-screen during analysis. An example is shown in Figure 5.8. The primary goal is to accurately determine the hotspot location.[2]

0. As an initial step, fiducials on the MOSS chip are defined once and are easily recognised by computer vision software using the OpenCV framework [153].

[2] While the temperature of the hotspot can be extracted by converting raw counts per pixel to temperature via a formula provided by the manufacturer of the thermal camera, the resulting number is only a rough approximation. The temperature measurement depends on the emissivity and reflectivity of the object. For example, even at thermal equilibrium, the MOSS sensor and its structure can be clearly distinguished from the gold-plated copper traces on the PCB and the PCB itself (see Figure 5.8 (a)), although all components are at the same temperature. As an estimate, in the most extreme case, a hotspot surface temperature of 10 °C above the surrounding MOSS temperature was observed.

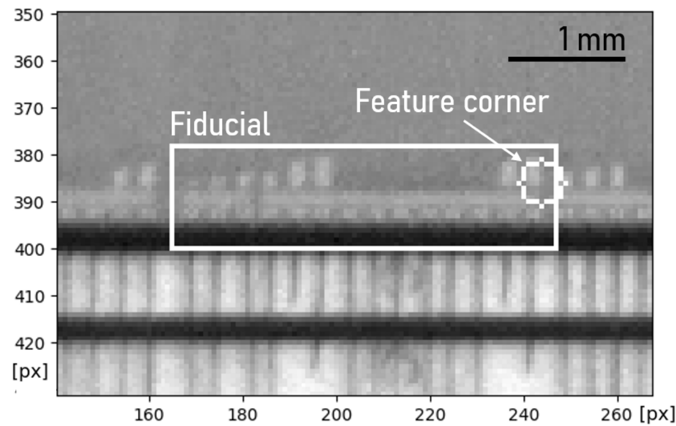


Figure 5.7: One fiducial area used to determine the chip coordinates from the thermal camera image. The reference feature corner is marked with a circle, and used to map the reference chip design coordinates and determine the transformation matrix. The black vertical lines in the bottom half of the image are wire bonds.

Eight locations are matched, and the locations of a corner of a metal structure is pre-determined in the fiducial image as illustrated in Figure 5.7.

1. After matching the fiducial images to the reference image under test (first image of the power ramp, before any power is applied), an affine transformation matrix is estimated using a least squares fit of the fiducial positions to the target positions of the global MOSS coordinate system. This step additionally allows the creation of a Region of Interest (ROI), considering only the MOSS sensor and excluding surrounding PCB structures in the next analysis steps. From the transformation parameters, the thermal camera pixel equivalent is estimated to be approximately 42 μm x 42 μm. Taking into account a slight optical barrel distortion at the edges of the image of about 1 pixel, a hotspot localisation accuracy of $\lesssim$100 μm can be expected in both directions. For later analysis, a best-case resolution of 50 μm × 50 μm and an average-case resolution of 100 μm × 100 μm are used.

2. Subsequently, all images are scanned by creating the absolute difference of the first image (no power applied) and each remaining image (powered at different levels). Only the previously determined ROI is considered in the next steps. For each of the n difference images, a empirically found global threshold is applied,

setting values below this threshold grayscale value to 0, followed by a median blur operation[3] with a 3×3 mask, reducing salt-and-pepper noise[4]. From these new images $n \cdot I_{ROI}$, the averages $n \cdot \bar{I}_{ROI}$ are calculated, and subtracted from the average value of a 5×5 pixel mask $\overline{M}_{5\times 5}(x, y, I_{ROI})$ which is scanned over the image. Looping over all n images the same way, the image which maximises the difference between the full ROI average and the 5×5 pixel mask average is taken as a candidate I_{cand} for further hotspot analysis:

$$I_{\text{cand}} = \underset{I}{\operatorname{argmax}} \left(\max_{x,y \in ROI} \left| \bar{I}_{ROI} - \overline{M}_{5\times 5}(x, y, I_{ROI}) \right| \right)$$

A selection on which part of the power ramp-up (which power net) to analyse can be made.

3. To visually highlight the hotspot location, a non-local means denoising step[5] is performed on the selected image (after applying difference, thresholding, and median blur). This allows to determine the contour containing the hotspot, and the coordinates of the maximum are extracted. This step is crucial for identifying hotspots which are very faint and not visible in the noisy difference or original image. During the semi-automated analysis, a user confirms the hotspots suggested by the analysis software. The centre-of-gravity is calculated as suggestion if multiple pixels exhibit the peak value. Cases of multiple simultaneous hotspots exist (cf. Figure 7.3).

4. The transformation determined in step (1) is applied to the operator-selected hotspot locations, and the coordinates in the MOSS chip-design coordinate system are stored.

[3] In a median blur of median filter operation, a mask (or kernel) is scanned over the image, and the median of the kernel area replaces the central pixel value of the kernel [154].

[4] Random single pixels of the thermal camera sensor exhibit fluctuating overexposure and/or underexposure. This effect is called salt-and-pepper noise and is effectively reduced by a median blur operation [154].

[5] Non-local means denoising considers an extended neighbourhood of each pixel to be denoised. The mean of the most similar pixels replaces the original pixel value. Fine structures are preserved [155].

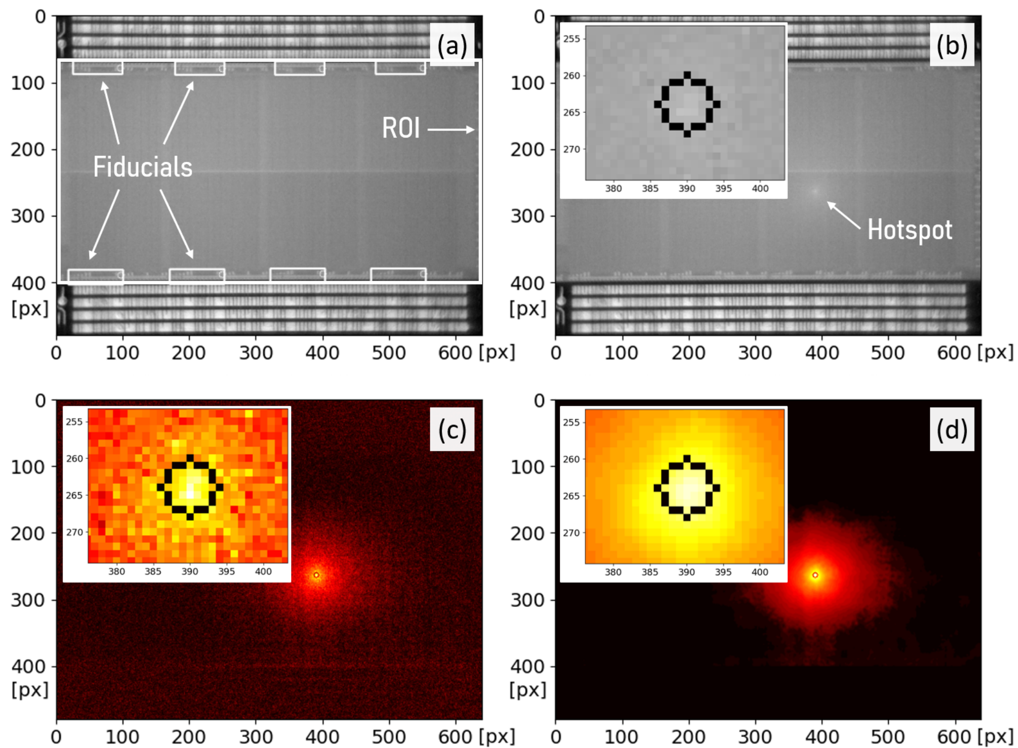


Figure 5.8: Thermal camera images and hotspot localisation: (a) The initial state of the HU prior to supplying power. This is the reference image. The white boxes indicate the fit of the fiducials used to calculate the transformation parameters to the global MOSS coordinate system. (b) The image containing the hotspot, found with the procedure described in the text. (c) Absolute, and (d) denoised difference between the reference image and the image with the hotspot. The location of the hotspot is shown magnified in the inset. For faint hotspots, denoising becomes crucial to distinguish the hotspot from noise (see also Figure 7.7).

5.2.2 Chip design correlation

The stored hotspot locations in chip design coordinates are subsequently overlaid with the chip design. Using KLayout [156], the area around the hotspot location is extracted from the chip design files with a custom script for further inspection. Additionally, a correlation with the impedance measurement is made. Power nets affected by a short (in metal layers M7 and M8) are highlighted in colour for visual representation as shown in Figure 5.9. Metals M7 are always routed horizontally, while metals in layer M8 are routed vertically (for the orientation shown here, where the left endcap of the chip is on the left, cf. Figures 3.14b and 3.15b). The best and average case resolution windows, as expected from the thermal camera imaging, are overlaid at the extracted hotspot location. The analysis then checks whether both power nets involved in the

short are present within either (or both) resolution windows and whether they overlap. This extraction is done automatically with custom software, and the illustrations, such as those shown in Figure 5.9, are saved for reference purposes only.

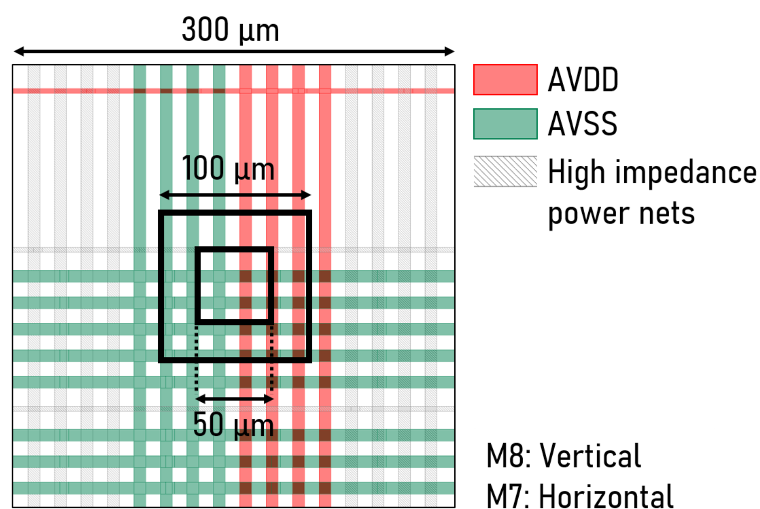


Figure 5.9: Example of a hotspot location mapping to the chip design (top-down view): The shorted power net pair is highlighted in red (AVDD) and green (AVSS). The M7 and M8 layer metals are routed horizontally and vertically, respectively. The extracted hotspot location is indicated by the central squares for the best, and average resolution windows, respectively.

5.3 Test system for functional tests

To functionally characterise the MOSS sensor, a dedicated test system (see Figure 5.10d) – referred to as the 'functional test system' – was developed. The MOSS sensor, mounted on the carrier PCB, is connected via 5 high-density 560-pin connectors to the so-called Proximity boards and 5 FPGA boards. One out of five FPGA–Proximity–MOSS blocks is shown as a diagram in Figure 5.10a. The readout PC communicates with the commercially available FPGA boards [157] housing Arria10 SoC (System on Chip) modules [158] via USB3. A custom firmware for the FX3 chipset (for USB3 communication) and integrated on the Arria10 SoC module is used. Each FPGA board is connected to a Proximity board. Each of the four Proximity boards (excluding the fifth board for the LEC) interfaces with each 5 HUs of the MOSS sensor via the long edge pads individually. The custom-designed Proximity boards have the following functionality (for each HU individually, unless otherwise noted):

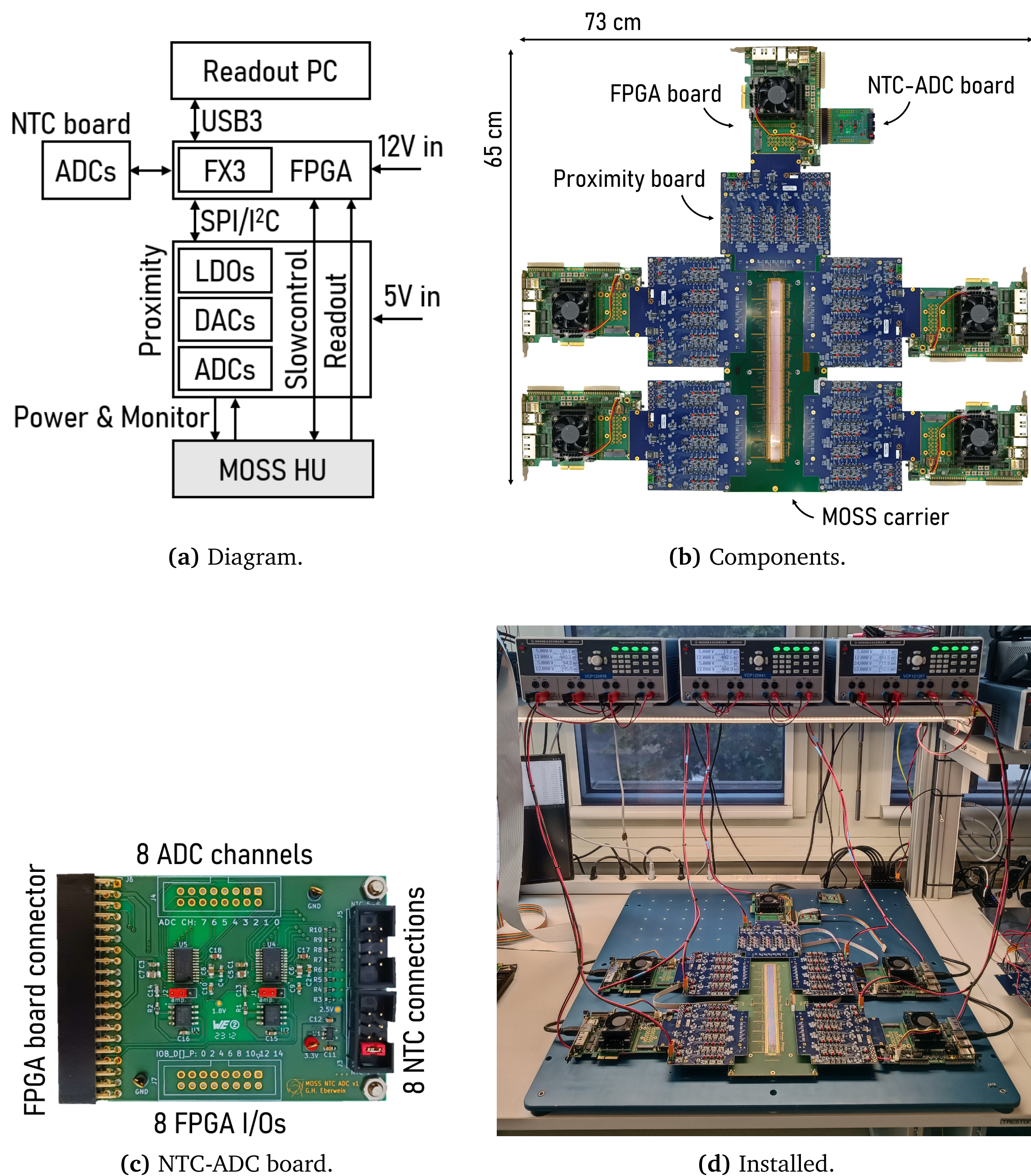
- Slow-control communication: Pass through (serial) write and read slow-control commands between the FPGA and the MOSS chip.
- Data readout: Pass through 8-bit parallel (hit) data from the chip for readout.
- Power delivery: Supply power to the chip using variable LDOs (Low-Dropout Regulators), which allow switching of each power domain individually. Supply voltages are set and controlled via DACs (Digital-to-Analog Converters), interfaced over I^2C through an I^2C channel multiplexer on the Proximity board. All power domains except PSUB are controlled in this way.
- Current monitoring: Measure current consumption individually for each power domain using shunt resistors ($O(0.5\,\Omega)$) and precision current sense amplifiers. The resulting voltage is digitised using ADCs (Analog-to-Digital Converters), interfaced via SPI. The measurement range is tuned to each power domain.
- Analogue monitoring: Read analogue output voltages from the MOSS chip's monitoring pads.
- Override capability: Provide voltages and currents to the MOSS chip's override pads via dedicated DACs. These can override the chip's internal DACs in the event of a malfunction.
- PSUB connection: The global (optional) PSUB voltage for the MOSS chip is supplied via a LEMO connector on any one of the Proximity boards connected to the carrier PCB. A separate power supply is used for this.
- Automatic identification: Each Proximity board and MOSS chip carrier PCB includes a unique ID chip, readable via the I^2C interface (also routed through the I^2C multiplexer). This enables the automatic identification of components in the software framework and the application of configuration parameters from a reference file.

- Interlock system: All MOSS power domains across all HUs are continuously monitored. If any current exceeds its limit on any FPGA–Proximity–HU chain, an interlock signal is generated and propagated via LEMO cables to all connected Proximity boards and FPGAs, shutting down all MOSS HUs in operation.

- Trigger and busy signalling: LEMO connectors are available to receive an external trigger signal or to issue a busy signal for triggered readout. An alternative trigger mechanism was adopted later for convenience (see below).

On a fully equipped test system with five Proximity boards, a total of 256 ADC channels, 520 DAC channels, and 100 voltage regulators are controlled. Following initial characterisation, several modifications to the Proximity board were necessary. These included adjustments to the physical pinout on the high-density connector to the FPGA, changes to jumper routings, and bug fixes related to the I^2C switch. Additional updates were made to the voltage divider values used for driving LDO outputs, as well as to the shunt resistor and amplification factors of the current sensing amplifiers, to better match the actual current consumption of the MOSS chip. A set of measurements of the DACs is shown in Figure 5.11a and Figure 5.11b for the voltage mode (12-bit resolution) and current mode (8-bit resolution), respectively. The output range can be set via I^2C commands and was chosen to match the requirements of the test system, with the shown measurements spanning the required range. All measurements are well within the specifications of the DACs. Measurements were performed with a $7\frac{1}{2}$-digit multimeter [150]. Similar tests were performed on the 12-bit ADCs, together with a precision voltage source [147], to confirm the specifications and readout modes. Understanding was gained in the operation of ADCs, DACs, and LDOs, including proper startup, configuration, and stabilisation time (e.g. needed for reference capacitors to fully charge, where measurements are invalid until a stable condition is reached). This was critical for the stable operation and monitoring of the MOSS sensor with the functional test system.

An additional NTC-ADC add-on board (see Figure 5.10c) was developed, allowing the measurement of the temperature of the MOSS sensor via NTC probes [159]



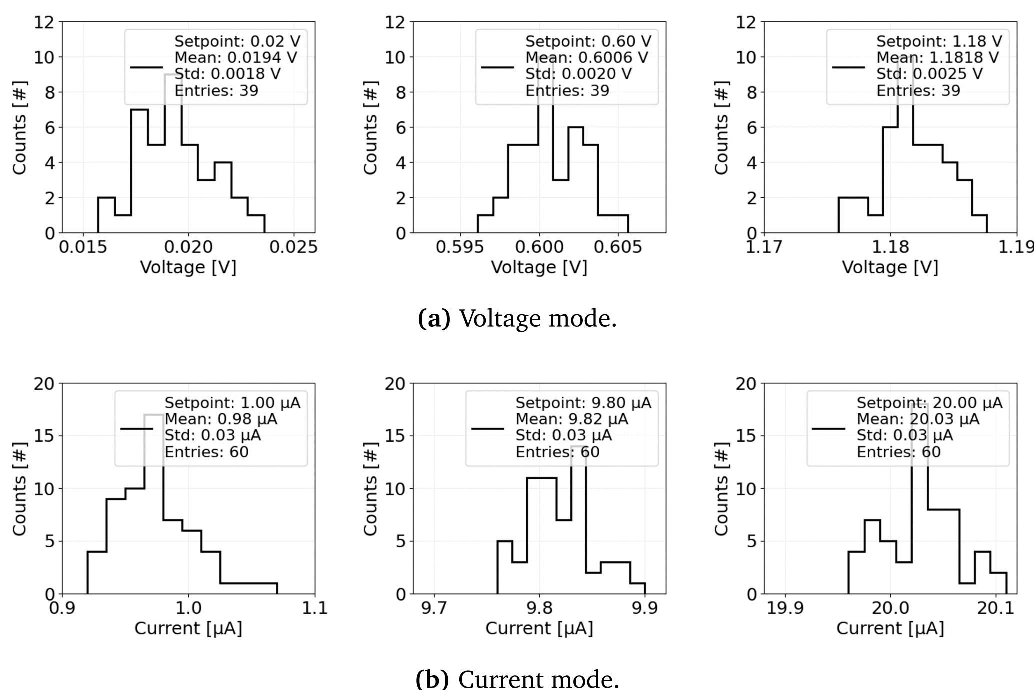
(a) Diagram.

(b) Components.

(c) NTC-ADC board.

(d) Installed.

Figure 5.10: (a) Schematic diagram of one out of five FPGA–Proximity–MOSS interfaces. (b) All boards required for a full test system. The board at the top connects to the LEC of the chip via one 560-pin connector on the MOSS carrier board. The other four FPGA–Proximity boards interface with each 5 HUs via each one 560-pin connector. (c) The NTC-ADC board. One of these boards is connected to (any) FPGA board and allows to measure the MOSS (board) temperature at 8 locations, and 8 additional voltages. Eight FPGA I/O channels are additionally available. (d) Fully assembled and installed test system in the laboratory with the protective cover of the MOSS chip removed. Three power supplies are visible at the top, with 10 individual channels required to power the full test system. The anodised blue aluminium base plate was custom-made, allowing for fast re-configuration and re-plugging of MOSS carrier boards.



(a) Voltage mode.

(b) Current mode.

Figure 5.11: DAC output voltages (a) and currents (b) for the ranges required in the MOSS test system.

mounted closely to the chip on the testing PCB. ADCs similar to those used on the Proximity boards were selected for this purpose. The voltage drop across each NTC, connected in series with a precision resistor (0.01% tolerance), is measured and converted to temperature according to the NTC specification. The reference voltage is generated on the board. The circuit diagram is given in the Appendix A.3. Two 8-channel ADCs are used, one for measuring the 8 NTC temperature probes mounted on the MOSS carrier PCB, and a second one allowing for probing up to 8 additional voltages as required. The board is directly plugged into an I/O-connector of any FPGA board used in the test system. The NTC-ADC board additionally passes through 8 FPGA I/O pins, allowing the connection of any logic signals, which was ultimately used to add a level-shifter board required to supply an external trigger signal used in test-beam measurements.

The test system software development started with characterisation and control of the Proximity boards. The test system software framework was then reworked and written as a Python Application Programming Interface (API), based on which the MOSS characterisation tests and scans are coded. The basic structure of the API is given

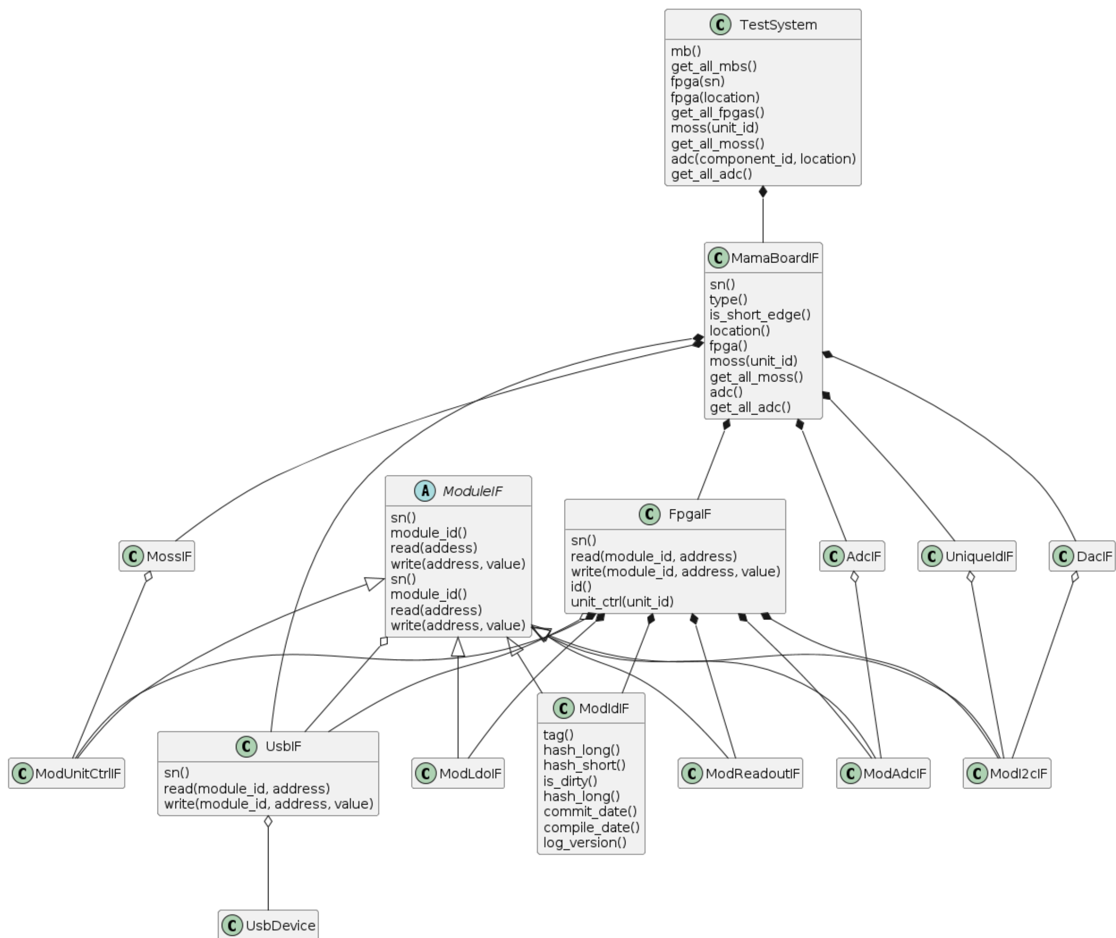


Figure 5.12: Diagram of the test system API written in Python illustrating the class and interface structure. Scans and functional chip tests are built upon this API.

in Figure 5.12. Custom FPGA firmware was developed within the group. It is worth noting that full operation of the test system and MOSS chip is not strictly required for testing. A single quarter of the chip (e.g., HU1–5) can be operated independently by connecting just one FPGA–Proximity pair to a corresponding high-density connector on the MOSS carrier PCB (e.g., the top-left connector).

Exhaustive measurements, from powering to single pixel characterisation, are performed with the functional test system, as outlined below. Tests are performed on HUs selected in a configuration file, where operational and test parameters are defined. Each test, or ‘scan’, creates a result directory in the database for a given MOSS sensor, where measurement, configuration parameter data, and log files are stored. Discussed here are scans performed from the long edge of the chip, excluding

tests of the LEC[6]. The following scans are used for sensor characterisation and are typically executed in the sequence outlined below:

- Test system power-up: Prior to running any scans, the test system itself needs to be powered up. The MOSS chip is placed into the test system, and the Proximity boards are connected to the MOSS carrier and FPGA boards. Then, the FPGA boards are powered, and the firmware is uploaded onto the FPGAs. Following successful FPGA programming, the Proximity boards are powered up. The test system is now ready for chip characterisation.

- Power-on scan: The power-on scan consists of two stages. First, each HU is powered up individually by supplying nominal AVDD, DVDD, and IOVDD voltages, providing a 33.33 MHz clock, toggling the reset signal, starting up the chip-internal DACs (specifically the bandgap references), and configuring the chip by writing to the register bank. Currents are measured on each power net at each power-up stage. If the HU is started up successfully, it is powered down again, and the next HU is powered up the same way. This sequence is repeated for every HU to be tested. The second stage of the power-on scan follows if every HU individually powers up correctly, staying within current limits. Each HU is then powered on sequentially, keeping the previous HU powered as the next one is additionally powered up. This is repeated until the full MOSS chip (or all HUs selected in the initial configuration file) is powered up. Currents are again measured and written to the results file. If successful, and currents are within limits, the chip is then kept powered on, and is ready for the next scan stage. Note: Initially, tests were performed with PSUB at 0 V. For tests with PSUB at –1.2 V, the voltage is applied prior to the first power-on scan, immediately after programming the FPGAs and powering up the Proximity boards.

[6] Adapted versions of the scans outlined here exist for the operation of the MOSS chip from the LEC and transferring data over the stitched backbone. These results are not discussed here, however, excellent agreement in chip configuration and readout was found when operating the chip from the LEC [7]. In this configuration, the individual HUs are only powered from the long edge as discussed in Section 3.5.1, with the full chip slow-control and readout from the LEC and across the backbone structures.

- Register scan: This scan verifies the integrity of the MOSS chip's register map and the functionality of the slow-control communication. The scan begins by reading the expected default values of the registers immediately after power-up, confirming proper chip initialisation. Then, a set of predefined values and patterns – along with one randomly selected value – are written to the registers and read back to verify correct communication and register behaviour. The scan concludes by restoring the original reset values through write and readback operations. In total, 340 out of 402 register addresses are tested. The remaining 62 registers are intentionally excluded, as they control on-chip DACs (which could put the analogue frontend into undesired states) and pixel mask configuration. These functionalities are covered in dedicated scans described below.

- Shift-register scan: Pixel masks (patterns, stored in registers tested previously) are written to the pixel matrices via shift registers. Similar to the register scan, patterns are now 'shifted in' and 'shifted out', again comparing written and read values to check for correct functionality.

- DAC scan: Each of the total 80 pixel matrices on a given MOSS chip is biased by a set of eight 8-bit DACs (four current DACs and four voltage DACs), totalling 640 DACs per sensor (cf. Section 3.5.1). Each DAC output is monitored via multiplexed access to dedicated output pads, with a linear output response expected. For current DACs, an external resistor is connected to the monitoring net and the effective current is calculated by measuring the voltage across the resistor. The voltage outputs of both current and voltage DACs are measured using the ADCs on the Proximity boards. Each group of four voltage DACs and four current DACs within a region shares a dedicated on-chip bandgap reference, which is also monitored via the multiplexed output pad. They are used to create on-chip references for the DACs, which are additionally measured via the monitoring pads. Trimming registers exist such that reference shifts are corrected during operation of the chip front-end. The reference and DAC output

values are written out in the results file. During the analysis, the reference spread, DAC minimum and maximum output, linearity onset and saturation onset, integral nonlinearity and differential nonlinearity, and fit slope and onset are extracted. The DAC outputs are always connected to the chip front-end of a given region. When testing one DAC, other DACs are put in a configuration such that the front-end stays as inactive as possible, minimising the effect of varying current consumption should a DAC be scanned through a region of otherwise high front-end activity. The currents on all power domains are stored during every step of the DAC scan.

- Digital scan: Each pixel on the MOSS chip includes integrated pulsing circuitry. This scan measures the pixel response to digital pulsing, enabling the first full exercise of the complete digital readout chain. Since the digital pulsing is injected behind the discriminator output of the front-end, this test isolates and verifies the digital data path, independent of the analogue front-end integrity. Malfunctioning pixels that fail to respond to the pulse are identified during this process.

- Analogue scan: Building on the digital scan, the analogue scan allows to additionally inject a predefined charge into each pixel front-end, now including the front-end in the pixel response test. Malfunctioning pixels are stored for masking, and readout issues, such as corrupted data packets, are logged. Pixels are classified by comparing the number of injections n_{inj} to the number of recorded hits n_{hit} per pixel (both for the digital and analogue scan): 'good' (n_{hit}=n_{inj}), 'noisy' ($n_{hit} > n_{inj}$), 'inefficient' ($n_{hit} < n_{inj}$), or 'dead' ($n_{hit} = 0$). Per default, $n_{inj} = 25$. One MOSS sensor has 6.72 million pixels.

- Fake-hit rate scan: The pixel front-end is set to a nominal working point by configuring the on-chip DACs. Each pixel matrix is read out $n = 100,000$ times, without external stimuli or pulsing. The number of hits per pixel is recorded and the fake hit rate (FHR) calculated for each region as $FHR =$

$\sum_{i=0}^{n=100,000} hits_i/(n \cdot pixels_{region})$. Pixels with a FHR over a defined threshold (default: 1%) are masked (typically, less than 3 pixels per region [7]).

- Threshold scan: The pixel front-end is again set to a nominal working point by configuring the on-chip DACs, at a defined threshold. The injection pulse (charge) is now varied, by varying one DAC voltage in a pre-defined range. Every (non-masked) pixel is scanned $n_{inj} = 25$ times. The resulting response allows to extract the threshold and noise for each pixel: Scanning through the injection pulse height range, the hits are read out and plotted, and a typical 'S-curve' or 'turn-on curve' is observed. The point of inflection is defined as the threshold. The derivative of the curve is a Gaussian with mean μ at the inflection point (threshold), and width σ representing the pixel noise. Threshold and noise are stored in DAC code units, and optionally converted to mV or e^- from a calibration measurement. For each pixel under test, the full matrix is read out, and pixels not under test, which still show hit counts, are masked as noisy and excluded from the threshold analysis. Finally, summary values per pixel matrix (region) are written to file, including RMS and mean of the threshold and pixel noise, and the number of noisy and bad (no hits) pixels.

- Source scan and testbeam measurements: For completeness, the source scan and testbeam measurements are mentioned here. A source scan is nothing else than a fake-hit rate scan with a radioactive source placed above the sensor, at varying distances and varying front-end parameters, randomly triggering the readout n_{trg} times. The testbeam measurement is done with a dedicated setup including trigger and tracking sensors, however, for the MOSS readout a FPGA–Proximity system is used together with the Python API to facilitate a triggered readout of the MOSS structures.

The very first chip-alive test was performed with the wafer probe card. The same FPGA–Proximity modules as introduced above were used; however, instead of connecting to the MOSS carrier PCB, the probe card was attached via the same high-density connector and extension cables, making contact with a single HU on the MOSS

chip through probe needles. Following a successful power-up, the first communication test was an attempt to write to an unallocated register address on the MOSS chip. This prompted the expected response – 0xDEAD – confirming communication with the chip and marking the beginning of the in-depth characterisation campaign [7].

6

MOSS Characterisation and Data Analysis

This chapter describes the measurement procedure, characterisation, and results, building on the test systems and measurements introduced in Chapter 5.

6.1 Measurement sequence

The measurement sequence for each MOSS chip is outlined below and shown in Figure 6.1. The data acquired are used for in-depth analyses discussed in the following sections.

1. Initial impedance measurement: Impedances of all 28 power net pair combinations per HU are measured. Shorts are classified as having an impedance of below 30 Ω.

2. Power ramping: First attempt to power each HU with nominal voltage. Each power net is individually ramped up in discrete voltage steps, and currents are measured. The ramp is stopped if individual current limits are surpassed. The power-up current shapes are stored. Transient high currents are extracted. The thermal camera is used to monitor the appearance of hotspots, allowing the extraction of fault locations. Each HU is classified in one of three categories.

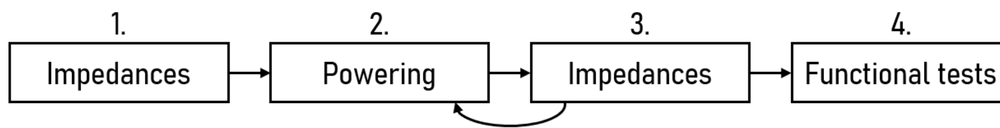


Figure 6.1: MOSS sensor measurement sequence.

3. Impedance measurement: After power ramping, the impedance of all 28 power net pair combinations per HU is re-measured. Changes in impedance between the initial impedance measurement and the repeated impedance measurement are extracted. A correlation with observed hotspots and transient high currents in the power ramp-up is made.

4. Functional tests: Functional tests are performed on HUs classified as testable in step 2, using the functional test system. Design structures on the MOSS chip are tested in detail, including power consumption, chip communication, DAC performance, pixel response, and noise behaviour as outlined in Section 5.3.

This sequence was defined and refined through multiple iterations of measurements using a set of approximately 10 wire-bonded and PCB-mounted MOSS sensors, initially supported by the wafer probing system. Thinning and dicing, and assembly of MOSS sensors on the carrier boards were a continuous process. Ultimately, 84 MOSS sensors, mounted and wire-bonded on carrier boards, were available for testing. Unless otherwise noted, a set of 80 MOSS sensors with comparable measurements will be discussed in the following sections. Four chips experienced (partial) physical damage, were used in destructive tests, or lack complete data for all measurement steps discussed. Where appropriate, wafer-probing data is additionally added to increase the data set size to 118 MOSS sensors. While first learnings on the chip performance were based on a subset of MOSS sensors, and MOSS sensors went through the measurement sequence at different times, summaries for all MOSS sensors tested are given where appropriate.

Measurement steps 2 and 3 were repeated multiple times. The power ramp-up sequence required several iterations to determine the optimal order and configuration for powering the chip. Additionally, current limits were gradually increased, enabling

functional testing of a larger fraction of HUs. This step-wise approach led to repeated power-up and impedance measurements, as discussed further below.

The following discussion will be largely based on the measurements up to the register scan, and the information they provide on yield and failure modes.

6.2 Impedance measurement results

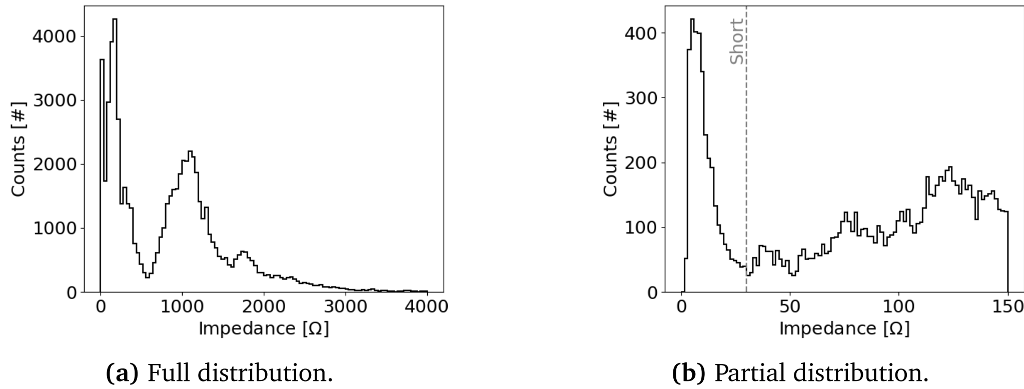
Results of the impedance measurements before powering the chip are discussed in this section.

6.2.1 Empiric definition of shorts at 30 Ω

The main goal of this measurement is to identify short circuit faults between power nets. Given the high complexity of the chip and parasitic diode structures, it is not feasible to extract expected, simulated impedance values from the chip design. Attempts to understand measured impedance distributions above values considered as shorted did not yield results. No comparisons between chip design and measured values in the non-shorted case are discussed. Further analysis of impedance distributions, however, has the potential to support failure analysis of silicon chips, especially when a reference distribution can be either simulated or measured with an appropriate number of samples.

The combined impedance distribution for all power net-pair combinations across 118 MOSS chips from 20 wafers is shown in Figure 6.2a (full range) and Figure 6.2b (up to 150 Ω). One sensor from wafer 17 was damaged, and one from wafer 24 was measured at a later stage; both are excluded from this dataset. These measurements were performed before power was applied to the chips. The chosen global cut at 30 Ω is indicated as a dashed line. This threshold was chosen based on the minimum between the first tail and the rise in the distribution. Diode structures – particularly on power nets spanning the full sensor – were simulated to yield low measured impedances, on the order of 50 Ω [160]. A measured impedance of below 30 Ω is called a short. The observed distributions of impedances for each of the $\binom{8}{2} = 28$

power net pair combinations individually are given in the Appendix A.4 for reference (split into top and bottom HUs).



(a) Full distribution.

(b) Partial distribution.

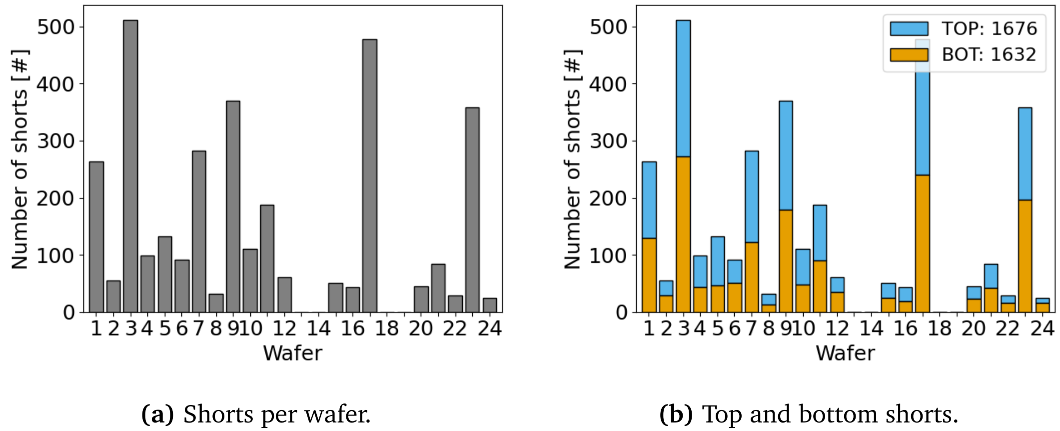
Figure 6.2: Distribution of impedances for all measured power net pair combinations of 118 MOSS chips: (a) Full distribution, and (b) low impedance region with cut at 30 Ω indicated as dashed line.

6.2.2 Shorts on each wafer

The number of measured shorts per wafer is shown in Figure 6.3a and Figure 6.3b overall, and split into contributions of the top and bottom HUs of the MOSS chips, respectively. The type of measurement (wire bonded on the carrier PCB, or wafer probed) for each wafer is shown in Table 6.1. Four out of 24 wafers produced were damaged during transport or thinning and dicing. Every wafer measured exhibits shorts. No MOSS sensor without shorts exists – every MOSS structure contains at least one HU with a short. The understanding and mitigation of the root cause of these failures is therefore crucial, as fabrication of ITS3 sensor layers relies on

Table 6.1: Measurements of impedances for all 24 wafers manufactured.

Wafer [#]	Measurement
1–8	probed, bonded
9–12	probed
13–14	none, broken
15–16	probed
17	bonded
18–19	none, broken
20–24	bonded



(a) Shorts per wafer.

(b) Top and bottom shorts.

Figure 6.3: Number of shorts measured for each wafer (a). The number of shorts in the top half and bottom half of the MOSS chip is similar (b).

high-yield sensor fabrication. A large fluctuation in the number of shorts between wafers is observed. This is compatible with a hypothesis of a processing issue rather than a systematic design fault.

A statistically significant difference in the number of observed shorts for odd-numbered and even-numbered wafers was observed, as shown in Figure 6.4. Two analyses were performed:

1. Point-biserial correlation: The point-biserial correlation coefficient r_{pb} is an appropriate choice if one variable is dichotomous, as is the case for the binary split in odd and even numbered wafers [161]. The split in groups of odd and even wafers (group 0 and group 1, respectively) is 50/50 with 10 wafers each. Mathematically equivalent to the Pearson correlation coefficient, the calculation simplifies to:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{6.1}$$

where:

$$s_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{6.2}$$

with:

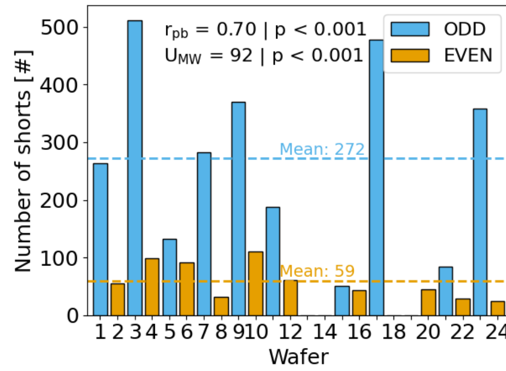


Figure 6.4: Number of observed shorts for odd and even numbered wafers with mean values indicated as horizontal dashed lines. Values for the point biserial correlation (r_{pb}) and Mann-Whitney U (U_{MW}) test are shown.

- M_0, M_1 the mean values of the continuous variable X for groups 0 and 1,
- n_0, n_1 the sample size of groups 0, 1,
- n the total sample size,
- s_n the standard deviation.

A strong correlation of $r_{pb} = 0.7$ with $p < 0.001$ is observed for the number of shorts split into odd and even numbered wafers.

2. Mann-Whitney U test: The Mann-Whitney U is a non-parametric test allowing to test for significant differences of two independent groups, determining if two samples come from the same population [162]. The U statistic is defined as the smaller of:

$$U_0 = n_0 n_1 + \frac{n_0(n_0+1)}{2} - R_0 \, , \; U_1 = n_0 n_1 + \frac{n_1(n_1+1)}{2} - R_1 \tag{6.3}$$

with:

- R_0, R_1 the sums of the ranks in group 0 and 1, respectively,
- $n_0 + n_1$ the largest rank after ranking all samples such that the smallest value obtains rank 1.

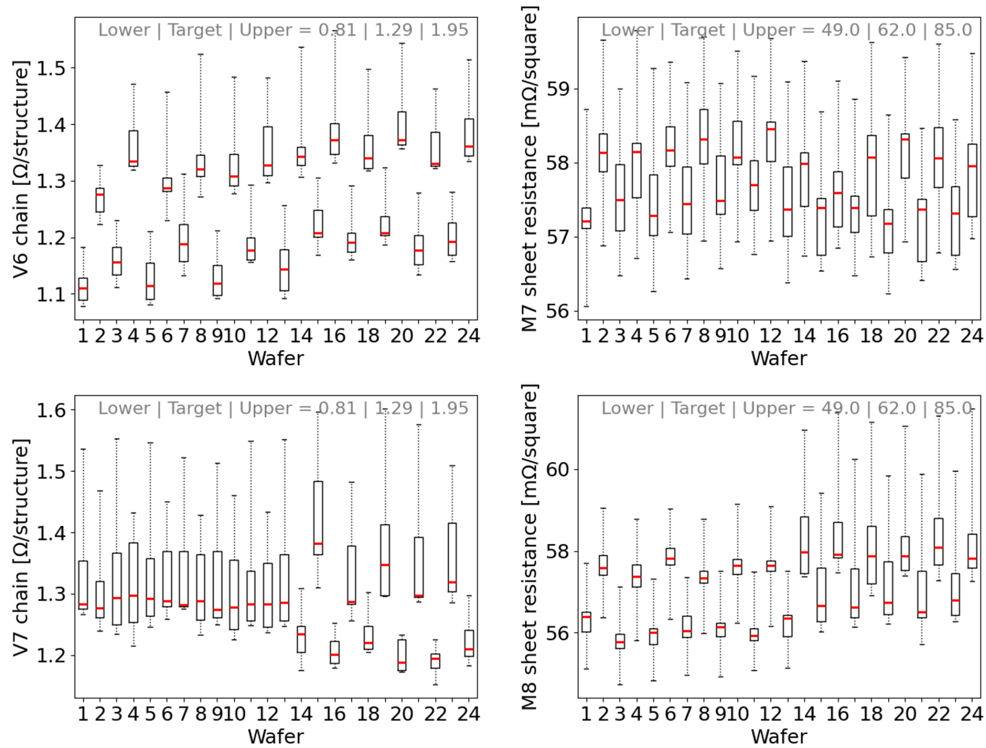


Figure 6.5: Wafer PCM data of sheet and via chain resistances of layers V6/M7 and V7/M8 for each wafer. Here, data for wafers 13, 14, 18, 19 exist, as these measurements were performed by the foundry and before wafer breakage.

With $n_0 = n_1 = 10$ (each 10 odd and even wafers) and $U = 92$, the straightforward interpretation is as follows: For $n_0 \cdot n_1 = 100$ pairwise comparisons, $U = 92$ are ranked lower from group 0 (even wafers) over group 1 (odd wafers), confirming a systematic difference between the groups with $p < 0.001$.

Wafer Process Control Monitoring (PCM) data was requested from the foundry. Around the perimeter of the central area of the wafer containing the 6 MOSS sensors, there are eight test structure sites incorporated by the foundry. The data indicate that all structures were within specification. However, a systematic pattern is observed looking at the data. For example, the resistances of the test structures in metal layers 7 and 8 (via chains V6, V7 and sheet resistances M7, M8) show clear trends in the measurements recorded by the foundry during quality control, as illustrated in Figure 6.5. Red lines indicate the mean values, the box spans from the first quartile to the third quartile, and the whiskers extend to the farthest data points on either side of the box. Fluctuating parameter values with odd and even wafer parity are

clearly visible for all shown datasets. An additional pattern in parameter values for wafers 1–13 compared to wafers 14–24 is visible for the V7 chain and M8 sheet resistance. It is plausible that, for example, odd and even wafers were processed in separate machines or lines, and wafers 1–13 and 14–24 at different times with slightly changing process conditions. This is one plausible hypothesis based on the data available. All parameters are within specification, and no definitive statement can be made regarding the cause.

When correlating PCM data (Figure 6.5) and the number of observed shorts measured for each wafer (Figure 6.4), distinct clusters for odd and even numbered wafers are observed as shown in Figure 6.6. The mean value of each wafer for the four parameters (V6, M7, V7, M8) is plotted against the corresponding number of observed shorts for each wafer. The point-biserial correlation r_{pb} is calculated for each parameter against odd or even wafer parity. All parameters show a strong (anti-)correlation with odd vs. even wafer numbering.

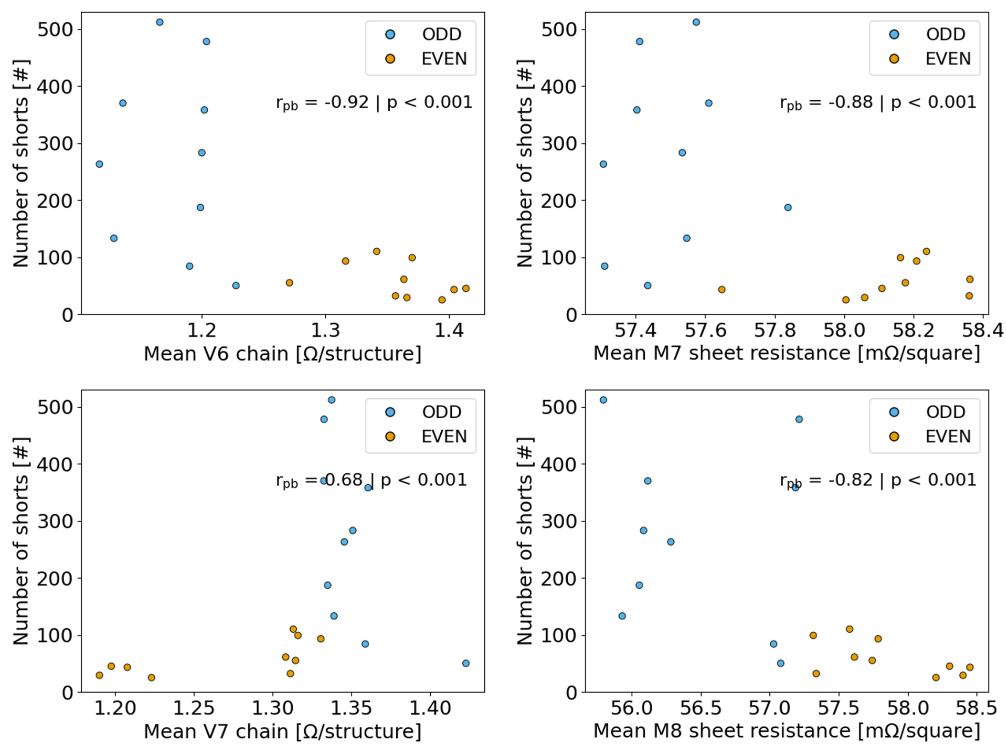


Figure 6.6: Correlation of number of shorts for each wafer with wafer PCM data for four parameters (V6/M7 and V7/M8). The point-biserial correlation coefficient is shown on each plot for the parameter values in dependence on odd or even wafer parity.

The number of observed shorts strongly depends on the measured wafer. A dependence on odd or even wafer numbering is confirmed. Wafer PCM data exhibits a similar trend, with all parameters within specification. Variations during processing between odd and even wafers are plausible.

6.2.3 Location of shorts on wafer level

The total number of observed shorts per HU for all measured wafers is shown on a wafer map in Figure 6.7. A larger density of shorts in the centre of the wafer is observed. Based on previous experience with MAPS fabrication (such as the ALPIDE chip), and following communication with the chip design and testing team, yield loss was rather expected at the edges of the wafer, in contrast to this observation. Many rotational symmetric manufacturing steps exist during wafer production. For example, photoresist spin coating, Chemical or Physical Vapour Deposition (CVD/PVD) of metals and dielectric, electroplating, Chemical Mechanical Polishing (CMP), and mechanical stress through wafer non-uniformity could lead to a central fault density gradient (cf. Section 3.4). Without knowledge of proprietary processing steps, it is

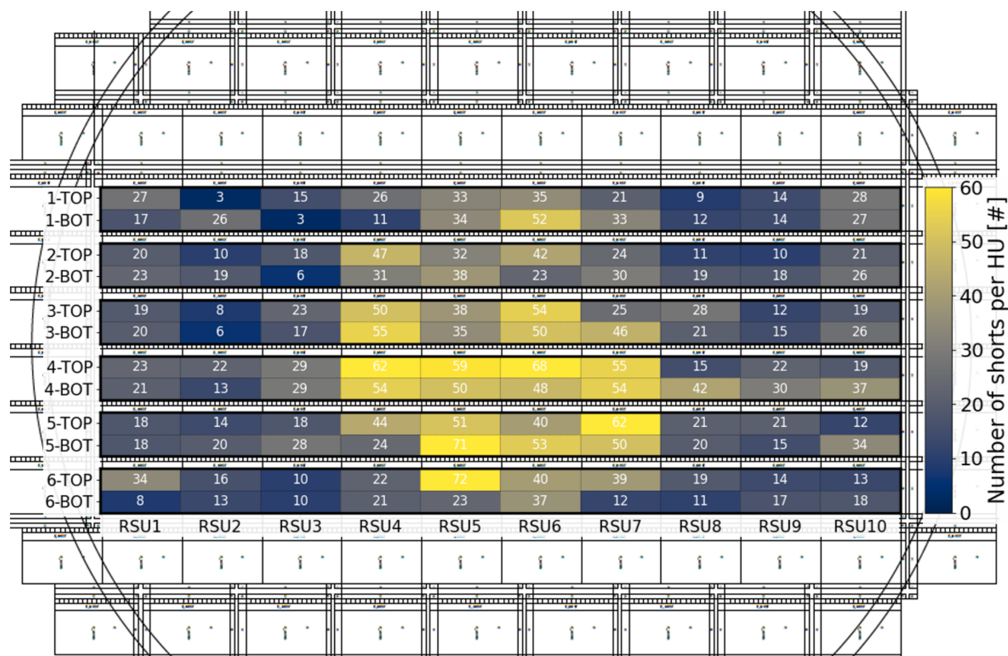


Figure 6.7: Total number of observed shorts for each HU for 20 wafers. An overlay of the data with the wafer shot map is shown. Each of the 20 HUs per MOSS sensor is shown, with 6 MOSS sensors per wafer split into top and bottom halves.

not possible to further narrow down the root cause at this stage. Wafer maps split into odd and even wafers are given in Appendix A.5.

6.2.4 Power net pairs affected by shorts

The distribution of the number of shorts for all $\binom{8}{2} = 28$ power net pair combinations is shown in Figure 6.8. No shorts are observed in power net pair combinations with BBVDD, BBVSS, or IOVDD nets[1]. Hence, only $\binom{5}{2} = 10$ power net pair combinations between power nets AVDD, AVSS, DVDD, DVSS, PSUB exhibit shorts. Furthermore, a lower number of shorts is visible for power net pair combinations involving the PSUB power nets. For the set of 10 power net pairs exhibiting shorts, there is no strongly favoured combination, supporting the hypothesis of a processing issue over a single systematic design fault. Analogous to Figure 6.3b, an even split in contribution to the number of shorts for each power net pair combination from top and bottom HUs is observed. The number of shorts is therefore not correlated with the low and high integration density of the top and bottom halves of the chip, respectively.

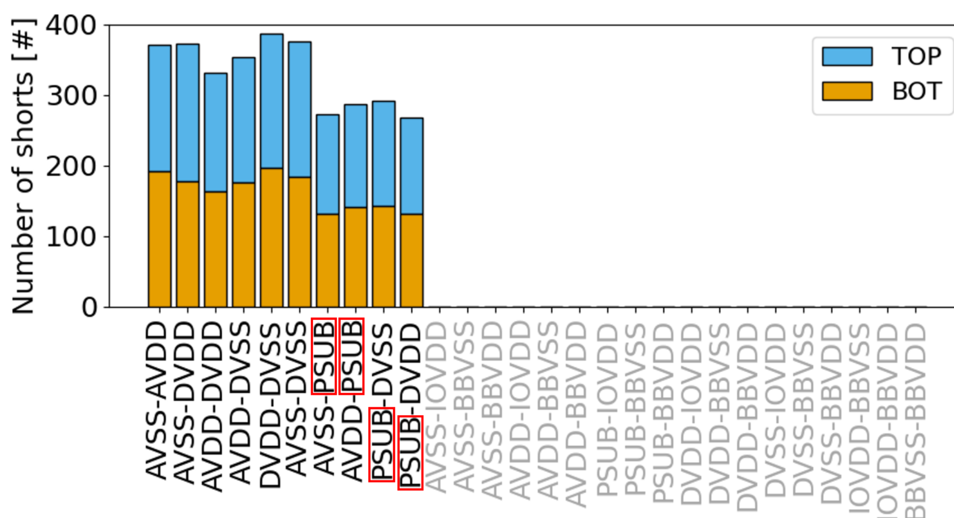


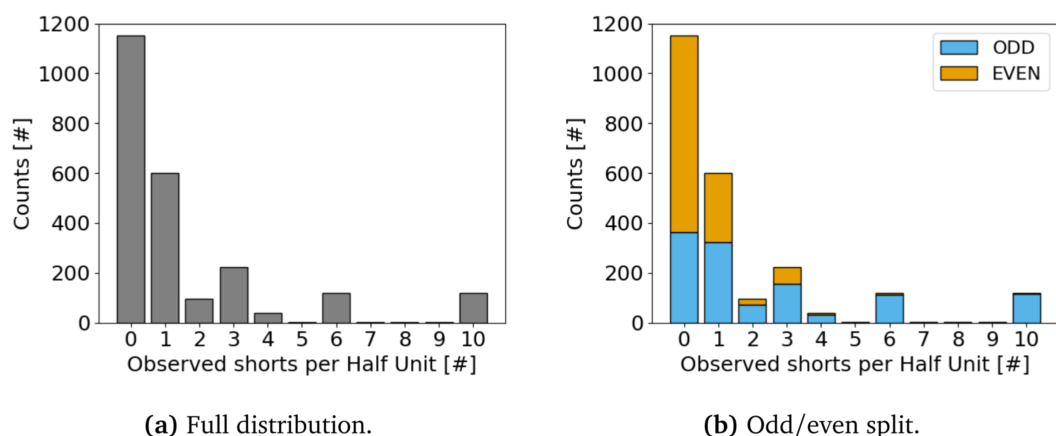
Figure 6.8: Number of observed shorts for all 28 power net pair combinations. Each bar is the sum of the contributions of the top and bottom HUs, showing an even split. No shorts are observed for power net pair combinations with BBVDD, BBVSS, and IOVDD power nets (grey). Power net combinations with PSUB show a slightly lower number of shorts (highlighted in red).

[1]On-chip routing differs for these power nets as further discussed in Section 7.2.

6.2.5 Observed vs. real number of shorts

HUs with more than one shorted power net pair are observed. The corresponding distribution of shorts per HU is shown in Figure 6.9a. If none of the 28 power net pair combinations for a given HU has a short, the observed number of shorts per HU is 0. The split in odd and even numbered wafers is shown in Figure 6.9b. Each 59 MOSS sensors for odd and even wafers were measured. For even-numbered wafers, 68% of measured HUs have no shorts, and 24% have one short. For odd-numbered wafers, 32% of measured HUs have no shorts, and 28% have one short (with the remaining HUs exhibiting more than one observed short).

It is appropriate to discuss the case of multiple observed shorts in more detail. When more than one short per HU is present, ambiguities and parasitically observed shorts exist. Consider, for example, three power net pair combinations: AVDD–DVDD, AVDD–AVSS, DVDD–AVSS, where two power net pairs are shorted. If now the third power net pair is measured, the shorted chain of the previous two power net pairs leads to an observed ('parasitic') short, although there is no physical short present. This is schematically shown on the left in Figure 6.10. Considering the 5 power nets and resulting 10 power net pair combinations, the graph representation in the centre and right of the image is used.



(a) Full distribution.

(b) Odd/even split.

Figure 6.9: Distribution of observed number of simultaneous shorts per HU for all wafers measured (a), and split into odd and even wafers in a stacked representation (b). The zero shorts bin represents the count of power net pairs with no observed shorts in the tested HU.

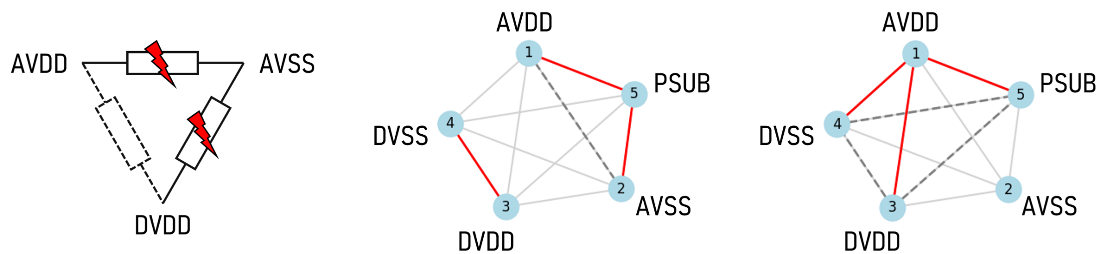
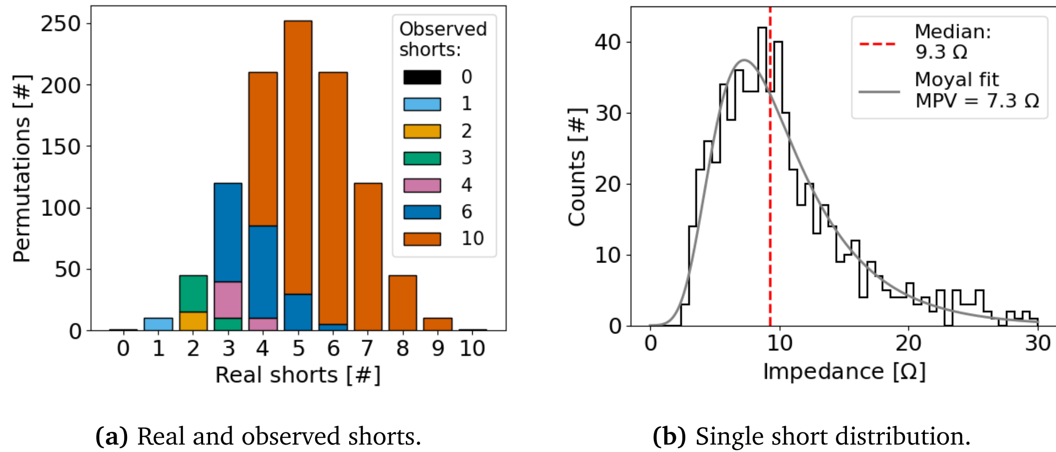


Figure 6.10: Schematic diagram of parasitic shorts for 3 power net pairs (left), and all 10 power net pairs (middle, right). On the left, two shorted net pairs (indicated by the red mark) lead to three observed shorts. The examples in the middle and right show configurations (and permutations thereof) with 3 physical shorts (red) leading to 4 and 6 observed shorts, respectively. The parasitic paths are indicated as dashed lines.

Using the graph representation and the NetworkX library for Python [163], all permutations of observed and parasitic shorts for 0–10 physical ('real') shorts were calculated. The resulting diagram – shown in Figure 6.11a – illustrates the number of real shorts, the corresponding permutations, and the number of observed shorts, with colour coding used to indicate different outcomes. For example, in the case of two physical shorts, the calculated permutations are 15 independent short pairs observed as two shorts. Additionally, 30 short pairs with a shared node result in three observed shorts. In the case of 4 real shorts, it is more likely to observe 10 or 6 shorts, since only 10/210 permutations lead to 4 observed shorts. Across all scenarios, only 0, 1, 2, 3, 4, 6, or 10 observed shorts occur – there are no valid configurations that produce 5, 7, 8, or 9 observed shorts. This is also visible in Figure 6.9. A negligible number of cases with 5, 7, 8, or 9 observed shorts appear, but these are attributed to the specific impedance threshold used in the analysis (see discussion below).

It must be taken into account that shorts exhibit non-zero impedance, and a chain of real shorts (i.e. a series connection) that leads to a parasitically observed short may exceed the 30 Ω threshold. The distribution of impedances with exactly one short is shown in Figure 6.11b. A fit with a Moyal distribution is performed to estimate the Most Probable Value (MPV) [164]. The longest chain creating a parasitically observed short is 4 real shorts in series. Given the shape of the short distribution, the most probable value of a short of $O(7\,\Omega)$, and median of $O(9\,\Omega)$, the cut-off for a short in Figures 6.9a and 6.9b is exceptionally set at 50 Ω to account for most parasitic shorts.



(a) Real and observed shorts.

(b) Single short distribution.

Figure 6.11: (a) Calculated permutations and observed shorts (binary) for pairwise shorts of 10 net pair combinations: 5, 7, 8, 9 observed shorts do not exist as calculated from the parasitic paths. (b) Distribution of impedances for shorts: Only HUs with one short are considered. The median and MPV from a Moyal fit [164] are shown on the graph.

It is now possible to run a simple simulation, attempting to match the distribution of Figure 6.9b. The following assumptions are made (including information from analyses performed later in this work):

- The underlying distributions of shorts for odd and even wafers differ.
- The probability for a short in a given HU (separately for odd and even wafers) follows a Poisson distribution. However, as shown in Figure 6.7, the spatial distribution of shorts across the wafer is non-uniform. To account for this non-uniformity, a composite distribution is introduced:

$$P(k;f,\lambda_1,\lambda_2) = f \cdot \frac{\lambda_1^k e^{-\lambda_1}}{k!} + (1-f) \cdot \frac{\lambda_2^k e^{-\lambda_2}}{k!} \tag{6.4}$$

  the sum of two weighted Poisson distributions with:

  - $f, (1-f)$ the weights of the first and second Poisson distribution,
  - λ_1, λ_2 the means of the first and second Poisson distributions,
  - k the number of shorts.

This model allows the simulation to reflect regions with differing defect densities (e.g., central vs. peripheral areas) and better reproduce the trends observed in the data.

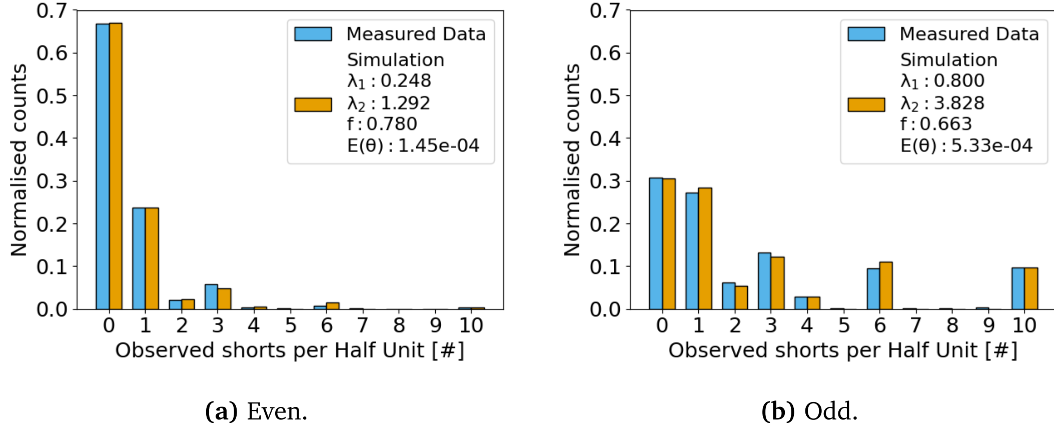
- The probability for each power net pair to have a short is not equal and therefore weighted. The four PSUB net combinations have a weight of 0.6, all other net pairs have a weight of 1.0 (see also Figure 6.8 and Figure 7.8d)
- It is possible to have more than one short affecting the same power net pair of the same HU. This is counted as one observed short, as it is not possible to discern from impedance measurements alone.

The simulation is then performed as follows, separately for odd and even wafers:

- Generate a set of 'real' short counts following the dual Poisson distribution for 0–10 shorts with each $n = 1,000,000$ tries.
- Generate the 'real' short distribution by picking random pairs of shorted power nets (weighted for PSUB net involvement), according to the set of simulated short counts.
- Generate the 'observed' short distribution, calculating all possible parasitic paths and avoiding double counts for > 1 short on a given net-pair.

A global fit is then performed separately for the observed shorts for odd and even wafers. The parameters λ_1, λ_2, f of the distribution defined in Equation 6.4 are varied to minimise the error, quantified as the sum of squared differences between the normalised simulated distribution and the normalised measured data. The measured data (target) are normalised as counts in bin $i = 1, ..., 10$ (observed shorts) over the sum of all observed shorts:

$$T_i = \frac{\text{count in bin } i}{\sum_{j=0}^{10} \text{count in bin } j} \tag{6.5}$$



(a) Even. **(b)** Odd.

Figure 6.12: Measured and simulated observed shorts per HU for even (a) and odd (b) wafers. Counts are normalised for comparison. The unweighted error $E(\theta)$ is quoted here.

Similarly, simulation counts, with parameters $\theta = (\lambda_1, \lambda_2, f)$, are normalised as:

$$S_i(\theta) = \frac{s_i(\theta)}{\sum_{j=0}^{10} s_j(\theta)} \tag{6.6}$$

The error is then defined as:

$$E(\theta) = \sum_{i=0}^{10} w_i \left[S_i(\theta) - T_i\right]^2 \tag{6.7}$$

with w_i optional weights (for example, to not consider the error on 5, 7, 8, 9 observed shorts in the measured data, or weighting each bin according to frequency assuming a Poisson distribution in each bin).

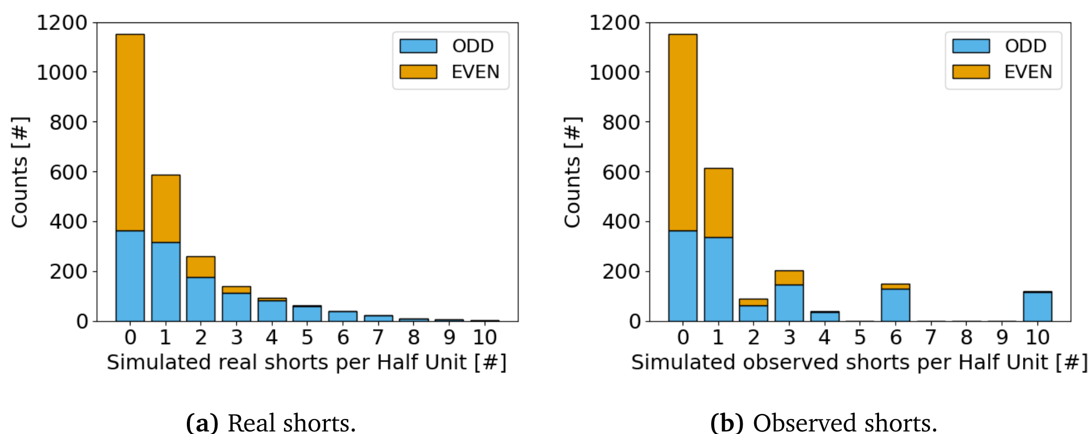
A simple grid search of all combinations Θ of varied parameters θ is then performed, finding the set of parameters $\theta^* = (\lambda_1^*, \lambda_2^*, f^*)$ which minimise the error $E(\theta)$ as:

$$\theta^* = \arg\min_{\theta \in \Theta} E(\theta). \tag{6.8}$$

Further error minimisation was performed with a differential evolution solver integrated in the Python SciPy framework [165, 166].

The resulting simulated and measured observed short distributions for odd and even wafers is shown in Figure 6.12, respectively.

The odd and even combined distribution of the number of simulated 'real' shorts (underlying distribution) and the resulting simulated observed shorts are shown



(a) Real shorts.

(b) Observed shorts.

Figure 6.13: Simulated real number of shorts, and corresponding distribution of observed shorts. The data are stacked for odd and even wafers, respectively. The distribution of simulated observed shorts closely matches the distribution of measured observed shorts (cf. Figure 6.9b).

in Figure 6.13, closely matching the measured distribution of observed shorts in Figure 6.9. For comparison, the simulated data was scaled to the same number of HUs without shorts as in the measured data.

A comparison of the number of total measured observed shorts and simulated real and observed shorts (scaled) is given in Table 6.2. This allows for the estimation of the fraction of real/observed shorts. Overall, approximately 75% of observed shorts correspond to actual physical shorts. After correcting for parasitic shorts, odd-numbered wafers exhibit a factor of ~ 3.8 more estimated real shorts than even-numbered wafers (without correction, the factor is 4.5 for observed shorts on odd vs. even-numbered wafers).

Note: To arrive at the total number of 'real' or 'observed' shorts, for > 1 short per HU, the count (number of occurrences of e.g. 2 shorts per HU) needs to be multiplied

Table 6.2: Comparison of total number of measured (meas.) and simulated (sim.), observed (obs.) and real shorts. The number of simulated shorts was normed to the number of observed shorts in the measurement for direct comparability. The number of real shorts can only be estimated from simulated data.

Dataset	Obs. (meas.)	Obs. (sim.)	Real (sim.)	Real/Obs. (sim.)
Odd [#]	2950	2962	2151	73%
Even [#]	649	674	564	84%
Combined [#]	3599	3636	2715	75%

by the number of 'real' or 'observed' shorts.

As before, and for the remainder of the text, no distinction is made between 'real' and 'observed' shorts, as the former can only be inferred through statistical estimation. The number of shorts is treated as the number of observed shorts. Consequently, the number of shorts per wafer, HU, and power net pair should be regarded as worst-case estimates, albeit unlikely.

Understanding and solving the short-circuit failure mode is crucial for successful fabrication of the final ITS3 sensor layers. At the observed failure density, a reliable ITS3 construction would not be feasible.

6.3 Powering results

The results of measurements with the power ramping setup will be discussed in this section. First, a classification of operational HUs will be made before analysing the power-on currents in more detail. Discussion of thermal camera images, analysis, and interpretation follows in Section 7.1.

6.3.1 Classification of Half Units and powering yield

First, powering of the MOSS chip is performed according to the method outlined in Section 5.2. The presence of shorts required an iterative process, increasing current limits stepwise and using functional tests to ensure the chip was not being affected. The first measurements were performed at PSUB = 0 V (ground potential), as the operation of the chip was foreseen in this configuration. However, at a later stage, it became clear that the chip needed to be operated at PSUB = −1.2 V to access its full performance[2].

Results here will be quoted with PSUB = 0 V and PSUB = −1.2 V. It was observed, and at that point understood, that shorts can (often) be opened by a 'burn-through', without harming the remaining chip. Additionally, after the root cause of the shorts was confidently established, the current limit was increased to 500 mA, attempting to open a maximum number of short faults. The current limits on the power nets used to

[2]Test beam measurements, especially after chip irradiation, revealed that for high detection efficiency at low fake-hit-rate, reverse bias is needed to speed up charge collection, reduce device capacitance and improve the S/N ratio [7].

test the chips in three systematic iterations of power ramps (ramp 1, ramp 2, ramp 3) are shown in Table 6.3. Nets with current limits of 500 mA always reach nominal voltage; however, a lower cut (operation current limit) is made for operation in the functional test system. A persistent high current above operational limits typically results in a persistent hotspot, visible with the thermal camera.

Note: A further increase in current, leading to potential additional burn-throughs, would only be possible by increasing the supply voltage, which would damage the chip.

To establish testing parameters, additional variations of the three power ramps discussed here were performed on subsets of chips, for both PSUB = 0 V and PSUB = −1.2 V. For safety, the operational limit of 25 mA on the PSUB net is, in case of ramp 3, applied after the initial PSUB power ramp-up. I.e. if during any of the subsequent power net ramp-ups, the 25 mA current limit on PSUB is surpassed, the power ramp is stopped and marked as unsuccessful.

Three classifications of power ramping outcomes are defined:

- OK-I: Power ramp-up of all power nets to nominal voltage, without transient high currents.

- OK-II: Power ramp-up of all power nets to nominal voltage, with transient high currents (burn-throughs). This corresponds to a jump in current (cf. Figure 5.6), and an impedance change from low to high on at least one power net pair. For easy classification across many iterations of power ramping measurements, a change of impedance on at least one power-net pair combination from low

Table 6.3: Power net current limits during power ramp measurements and for further operation of the chip in the full test system. The current limit on PSUB is only applicable if PSUB is powered.

Power net	Current limit (ramp 1) [mA]	Current limit (ramp 2) [mA]	Current limit (ramp 3) [mA]	Current limit (operation) [mA]
AVDD	100	500	500	100
DVDD	100	500	500	100
IOVDD	10	10	10	10
BBVDD	50	50	50	50
PSUB	n.a.	n.a.	500/25	n.a./25

($< 30\,\Omega$) to high ($\geq 50\,\Omega$) is used to identify a burn-through on a given HU. Alternatively, a relative impedance increase of more than 500% for a given power-net pair is considered an accurate indicator of a burn-through. Cases with more than one burn-through during power ramp-up for a given HU are observed. Burn-throughs are commonly accompanied by a disappearing hotspot visible with the thermal camera.

- LIMIT: Power ramp-up of all power nets to nominal voltage with persistent high current, or failure to ramp up all power nets within current limits. A persistent hotspot is commonly observed with the thermal camera, indicating an unrecoverable short.

A set of 82 and 80 sensors from 14 wafers (84 MOSS sensors in total, two broken during assembly) mounted on the testing PCB were measured, at PSUB $= 0\,\text{V}$ and PSUB $= -1.2\,\text{V}$, respectively. Two assembled and PCB-mounted MOSS sensors were partially damaged and/or used for destructive testing between power ramping measurements. Data are shown for the 80 MOSS sensors contained in three datasets for ease of comparison. The resulting distributions per wafer, and summary pie charts for odd and even wafers, are shown in Figure 6.14, Figure 6.15, and Figure 6.16 for ramp 1, ramp 2, and ramp 3, respectively. Comparing ramp 1, and ramp 2, it is evident that with a higher current limit, more HUs can be powered within operational limits. As a result, the fraction of sensors classified as 'OK-II' increases, while the fraction classified as 'LIMIT' decreases. HUs without any shorts always pass the power ramp and are categorised as 'OK-I', with 11 exceptions where a short has a value between 30–50 Ω (and the ramp is classified as 'OK-II'). Overall, 791/1600 (49%) HUs without shorts exist, all completing the power ramp-up. The larger fraction of 'OK-I' of 61% for ramp 1 and ramp 2 is attributed to the fact that shorts between ground nets or supply nets at the same potential do not contribute to a powering yield loss (see also Section 6.4).

Ramp 3 shows the results when applying PSUB $= -1.2\,\text{V}$. More ground net shorts (PSUB–AVSS, PSUB–DVSS) need to be opened by burn-throughs. These shorts do

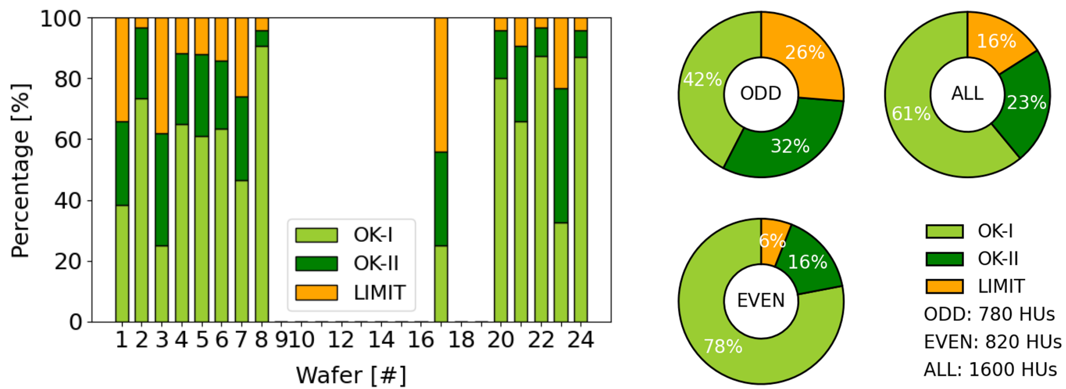


Figure 6.14: Ramp 1: Powering yield (80 sensors). PSUB = 0 V, AVDD/DVDD compliance limit: 100 mA. 'LIMIT' refers to HUs reaching the compliance limit.

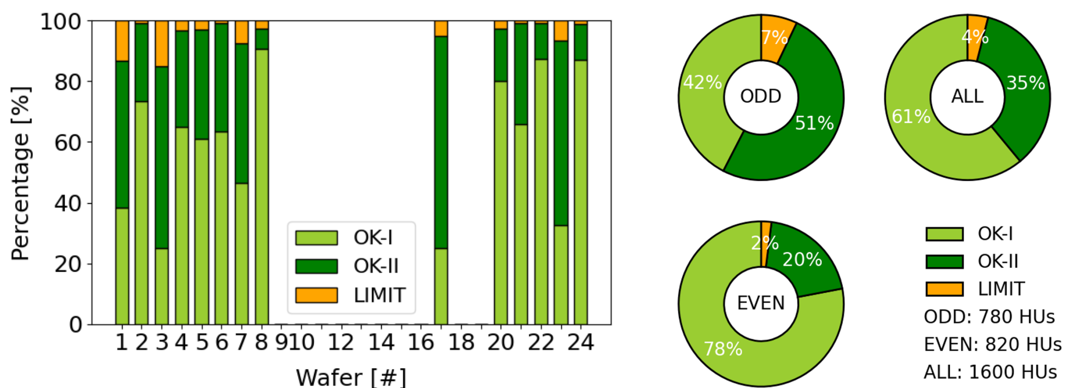


Figure 6.15: Ramp 2: Powering yield (80 sensors). PSUB = 0 V, AVDD/DVDD compliance limit: 500 mA. 'LIMIT' refers to HUs with currents above operational limits.

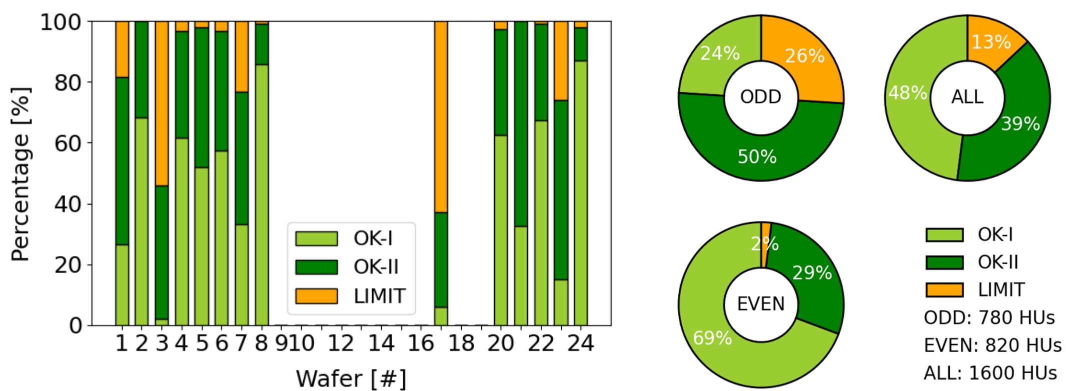


Figure 6.16: Ramp 3: Powering yield (80 sensors). PSUB = −1.2 V, AVDD/DVDD/PSUB compliance limit: 500 mA. 'LIMIT' refers to HUs with currents above operational limits.

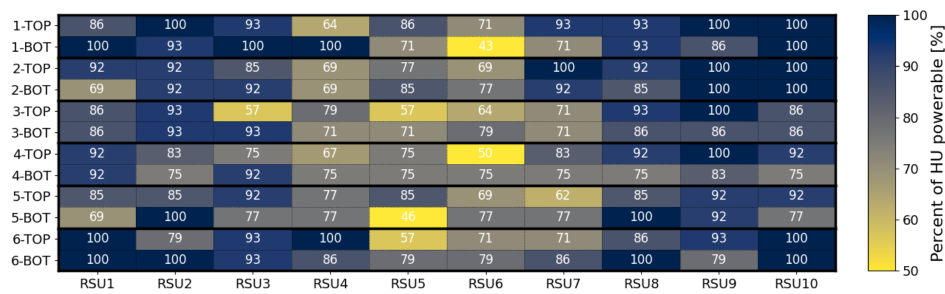
not affect powering at PSUB = 0 V, as the nets are on the same ground potential. The fraction of operational HUs stays the same for even-numbered wafers (with an increase in the fraction of OK-II classified HUs). Odd-numbered wafers show a significantly reduced operational fraction compared to PSUB = 0 V. The data show, that this fraction is dominated by wafer 3 and wafer 17, with > 50% HUs exceeding operational limits. More conservative operational limits on the PSUB power net (25 mA, see Table 6.3) lead to a reduction of operational HUs and an increase of the fraction of 'LIMIT'.

Generally, it is observed that repeated power ramping can lead to additional burn-throughs on all remaining shorted power-net combinations. Shorted power net pairs with AVDD or DVDD, therefore, are at least a factor of 3 more often attempted to be burnt through compared to PSUB to ground net shorts. In particular, MOSS sensors from wafers 17–24 underwent a higher number of initial iterations at PSUB = 0 V, using lower current compliance limits.

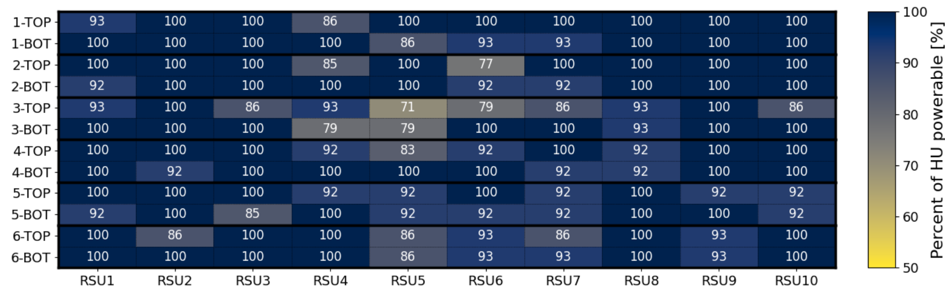
Powerable HUs on a wafer map level are shown in Figure 6.17 after ramp 1, ramps 2, and ramp 3, respectively. The quoted percentages refer to the number of powerable HUs relative to the number of MOSS sensors present per wafer position (1–6), accounting for any broken chips. A correlation between the distribution of powerable HUs and distribution of number of shorts (see Figure 6.7) is evident for ramp 1 and ramp 2. The increase in powerable HUs after ramp 2 highlights the impact of the higher current compliance, which leads to an increased number of successful burn-throughs.

The distribution after ramp 3 differs. This is driven by three effects: (1) Repeated power ramping on AVDD and DVDD power nets further increases the HUs within operational limits at PSUB = 0 V. (2) PSUB power nets are global nets that span the entire MOSS chip and connect across all HUs. Due to the operational current limit of 25 mA (see also Table 6.3), parasitic current paths both within individual HUs and across the full MOSS chip lead to non-powerable HUs. Each HU has separate AVSS and DVSS ground nets. A short between PSUB and individual ground nets, therefore, does not lead to a global short. Such a short contributes, however, to a global leakage

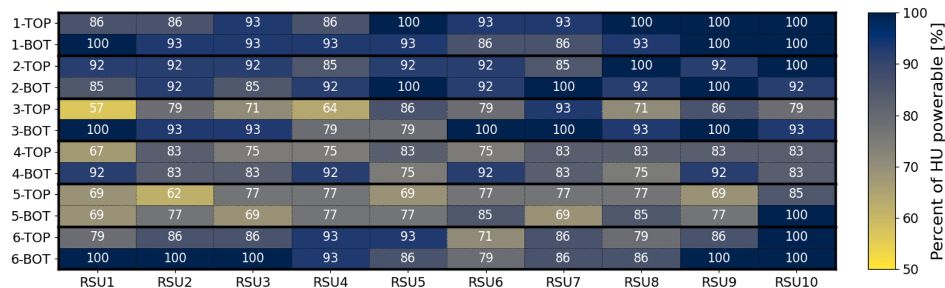
increase. (3) A coupling of PSUB and the global backbone nets (BBVDD/BBVSS) was observed (see Figure 6.18). This is understood to arise from protection and parasitic diode structures, whose presence and behaviour vary across sensors. This coupling leads to an increase in current on the PSUB power net during the ramp-up of the BBVDD net, occasionally exceeding the 25 mA operational current limit. It is worth noting that a more centrally peaked gradient in powerability is again observed when powering units within the functional test system, where BBVDD was not applied and the PSUB current limit was partially increased to 50 mA (see Appendix A.6).



(a) Ramp 1. PSUB = 0 V, AVDD/DVDD compliance limit: 100 mA.



(b) Ramp 2. PSUB = 0 V, AVDD/DVDD compliance limit: 500 mA.



(c) Ramp 3. PSUB = −1.2 V, AVDD/DVDD/PSUB compliance limit: 500 mA.

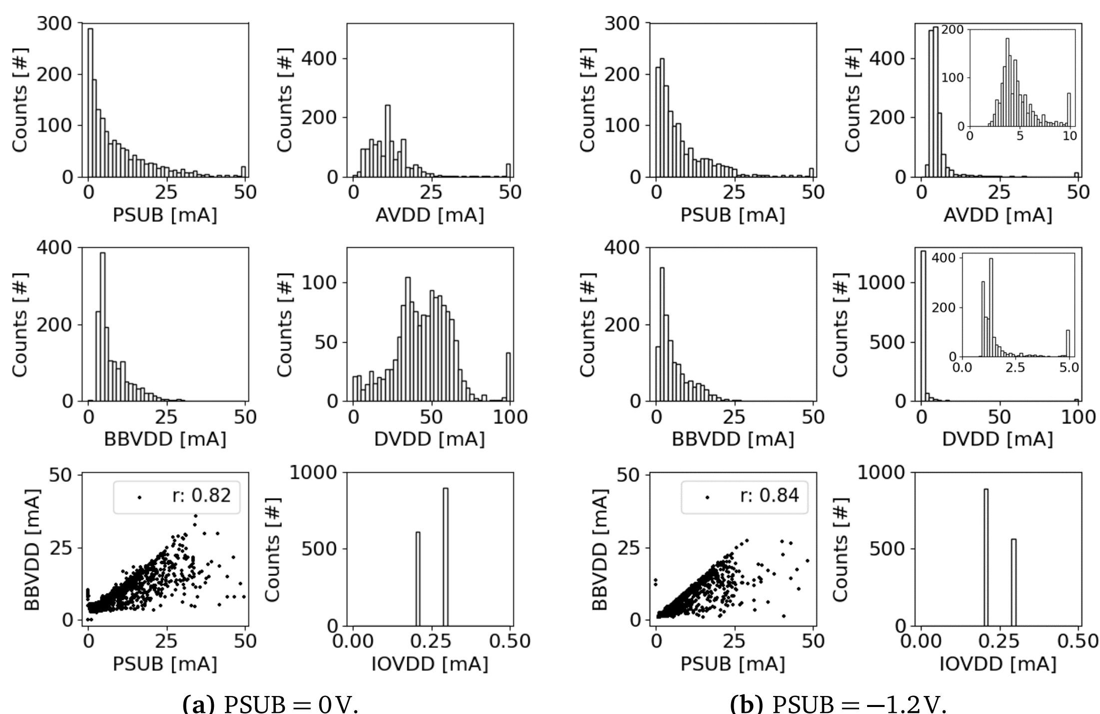
Figure 6.17: Powerable HUs per number of MOSS sensors (for each position 1–6) tested. Results after ramp 1, 2, and 3 are given in (a), (b), and (c), respectively.

6.3.2 Power-on endpoint currents

The currents in the powered-on state, completing the power-on ramp (ramp 2 at PSUB = 0V and ramp 3 at PSUB = −1.2V) are given in Figure 6.18a and Figure 6.18b, respectively. For the case of PSUB = 0V, the measured PSUB current is the measured return current between the AVSS/DVSS ground nets and the PSUB net (see Figure 5.5a). No difference in distributions was observed between the top and bottom HUs. Currents exceeding the operational limits are included for completeness; however, it is required that HUs successfully complete the full power ramp to be considered powerable.
Following observations are made (Figure 6.18):

- The PSUB current distribution, both at 0 V and −1.2 V has a peak at below O(5mA) with a smooth falling tail. The BBVDD current distribution closely matches the PSUB current distribution, and a strong correlation is observed (Pearson $r > 0.82, p \ll 0.001$). This correlation is attributed to a coupling between the backbone power nets and the chip substrate via protection diode structures and parasitic diode structures. Note that for PSUB = −1.2V, a cut-off in PSUB current is visible at 25 mA – the power ramp current limit after initial PSUB bring-up (where the limit is 500 mA). The few entries above 25 mA are due to misconfiguration of the measurement setup, not stopping the power ramp at 25 mA.

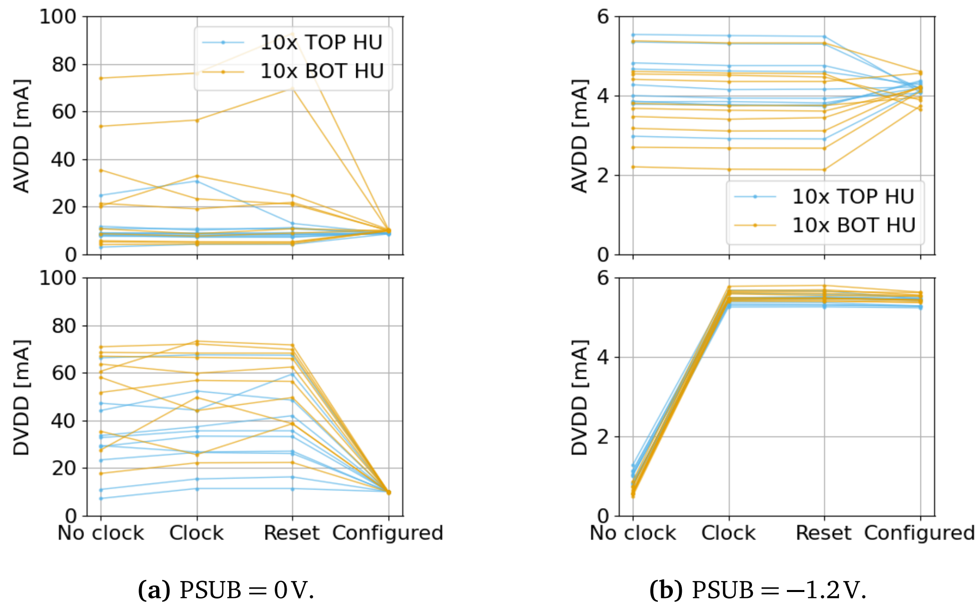
- AVDD and DVDD power nets exhibit a factor $O(3)$ and $O(50)$ wider distribution comparing PSUB = 0V and PSUB = −1.2V, respectively. This is a substantial effect, particularly for the DVDD current spread at PSUB = 0V, which was unexpectedly large. This behaviour is explained as follows: The chip's internal reset requires the setting of a register, which is not done automatically at power-up. Additionally, to start up the so-called bandgap references, which provide a reference voltage to the internal DACs, separate registers need to be toggled via the slow-control interface for each HU. If these configurations are not properly



(a) PSUB = 0 V. **(b)** PSUB = −1.2 V.

Figure 6.18: Endpoint currents of all HUs completing the power ramp-up (including outliers outside operational limits) for (a) PSUB = 0 V and (b) PSUB = −1.2 V. The last bin in each histogram is an overflow bin. For comparison, the axis limits are matched between (a) and (b). Therefore, an inset for both AVDD and DVDD at PSUB = −1.2 V is used for better representation of the current distribution.

set, the DACs provide non-controlled voltages and currents to the chip front-end, which in turn consumes a range of power given the ill-defined working point, leading to a spread in the AVDD currents. Since the front-end and the discriminator output are always active when AVDD is supplied, this leads to significant and uncontrolled digital activity, reflected in the large spread of DVDD currents. If the chip front-end is in an ill-defined working state, the discriminator output can be strongly active (it is not possible to switch off the chip front-end and/or discriminator output), translating to high activity on the digital domain, visible as very large spread of the DVDD currents. This behaviour is illustrated in Figure 6.19a. Using the functional test system, each HU is powered up and configured by (1.) applying power, (2.) supplying a clock signal, (3.) applying the reset, (4.) toggling the bandgap references and configuring the registers. Only after the last configuration step, the chip enters a well-defined operating state, with stable and uniform current draw in both



(a) PSUB = 0 V.

(b) PSUB = −1.2 V.

Figure 6.19: Power consumption on analogue and digital domains for one MOSS sensor (10 top and 10 bottom HUs) during the power-up sequence in the functional test system at PSUB = 0 V (a) and PSUB = −1.2 V (b).

the analogue and digital domains. During the power ramp-up measurement, it is not possible to configure the chip, and therefore, the endpoint currents carry limited information.

In the case of PSUB = −1.2 V, the chip behaviour is different. Applying PSUB changes the chip front-end response, resulting in reduced activity even when on-chip DAC settings are ill-defined. This is plausible, as each transistor PWELL and the chip substrate (PSUB) are on the same potential, effectively shifting transistor working points and the collection diode response. This leads to the aforementioned reduction in the spread of endpoint power-on currents after power ramp-up. Similarly, the power-up sequence with the functional test system changes (see Figure 6.19b): current consumption on the digital domain increases after supplying a clock signal and remains constant. The analogue domain currents still exhibit some spread given the ill-defined DAC states, and reduce after start-up and configuration of the DACs.

After power ramping without chip configuration, the AVDD currents typically range from 2 to 25 mA, and DVDD currents from 1 to 75 mA at PSUB = 0 V.

Currents up to O(100mA) are observed to fall into a well-defined state after configuration in the functional test system (cf. Figure 6.19a). At PSUB $= -1.2$ V, AVDD and DVDD currents are mostly well-behaved even without configuration, with ranges of 2 to 15 mA and 1 to 2.5 mA, respectively.

- IOVDD currents remain below 1 mA, with the measurement precision limited by the power supply resolution limit of 0.1 mA. Such a low current is expected, as the IOVDD domain, responsible for the signal level translation between 1.2 V (on-chip) and 1.8 V (off-chip), is inactive during initial power-up.

The power ramp measurements provide a first indication and an upper limit of the number of functional HUs on one MOSS sensor.

6.3.3 Burn-through current and voltage

The distribution of burn-through currents and voltages is discussed here. The last measurement point (voltage–current pair) before a burn-through is used as a metric. A simple threshold criterion was applied: a drop of more than 10 mA between two consecutive measurement steps (each with a fixed voltage increment of 100 mV) during the power ramp-up of any power net. Only data from HUs with at least one impedance change from low to high, with endpoint currents within operational limits, are considered. Data from ramp 1, ramp 2, ramp 3 are combined. Overall, 730 jumps are detected, and the resulting distributions are shown in Figure 6.20. Histograms are stacked, i.e. each bin indicates the PSUB, AVDD, and DVDD contribution to the number of bin entries. Burn-throughs are classified as AVDD if, during AVDD ramp-up, a short to AVSS, DVSS, or DVDD is opened. Similarly, burn-throughs are classified as DVDD if, during DVDD ramp-up, a short to AVSS, DVSS, or AVDD is opened. If PSUB is involved, the jump is classified accordingly – either if a short to PSUB occurs during AVDD or DVDD ramp-up, or, in the case of ramp 3, during PSUB ramp-up (note that –1.2 V is mapped to +1.2 V in this context) with a short to AVSS or DVSS. This chosen classification explains the seemingly larger PSUB contribution.

The discrete binning in the voltage histogram stems from the discrete voltage steps during power ramp-up.

A burn-through does not require a high current: 83.4% of measured burn-throughs appear at a current below 100 mA – the operational limit for AVDD and DVDD power nets. Less than 1.7% of burn-throughs require currents above 200 mA. These moderate currents do not pose any danger to the chip power supply network.

In conventional 'smoke tests' of silicon chips, the devices are typically immediately powered with the nominal supply voltage. Given these data, it is clear that a burn-through would therefore not have been detected, and the chip simply classified as 'OK', masking a potential underlying processing defect and reliability concern. The use of impedance measurements and slow power ramp-up are powerful tools for early fault detection.

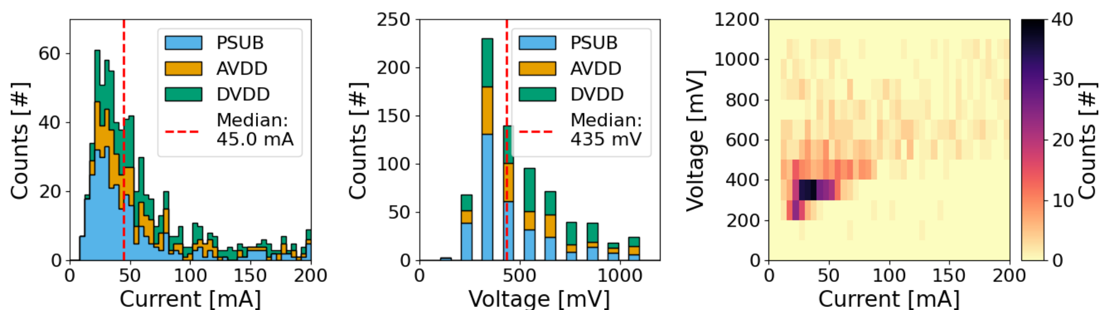


Figure 6.20: Distribution of burn through currents (left) and voltages (middle). Histograms are stacked, with each bin the sum of the PSUB, AVDD, and DVDD contributions. A 2D histogram correlating burn-through currents and voltages is shown on the right, with a distinct concentration around the maxima of both individual distributions.

6.4 Post power impedance measurement results

After power-ramping the chip, the impedances are again measured. Changes of impedances from low to high (short to no short) are observed. For comparison, this section focuses on the 80 MOSS sensors that were mounted on the carrier PCB and systematically evaluated using the power ramping setup.

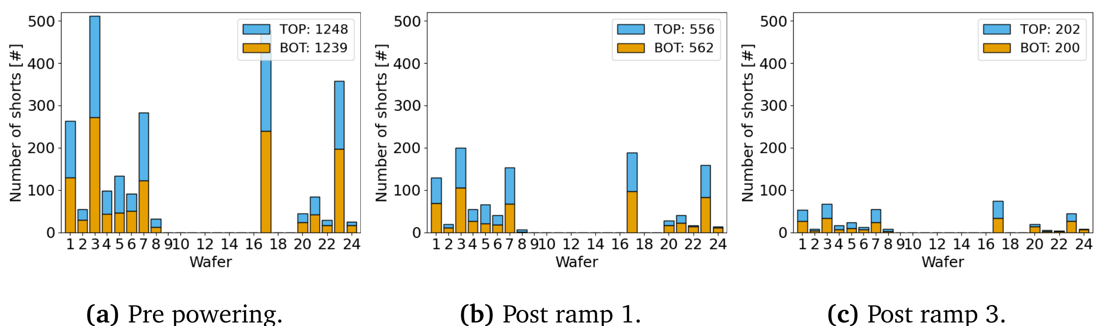
Impedance measurements taken before applying power (pre-powering), after power ramp 1 (post ramp 1), and after power ramp 3 (post ramp 3) are compared. Details of the ramping procedures are provided in Table 6.3.

In Figure 6.21, the number of observed shorts for each wafer, split into top and bottom HU contributions, is shown for the three measurement steps. The split of top and bottom HU contributions to the number of observed shorts is equal across all three datasets. With increasing current limits, and powering the PSUB net, the number of burn-throughs increases, leading to a reduction in observed shorts. After ramp 3, 84% of observed shorts are opened up.

The number of observed shorts per power net pair combination is shown in Figure 6.22 before powering, after ramp 1, and after ramp 3, respectively. The initial distribution (Figure 6.22a) matches the distribution discussed in Section 6.2.4. After ramp 1, a pattern emerges, and four power net pair combinations need to be discussed further:

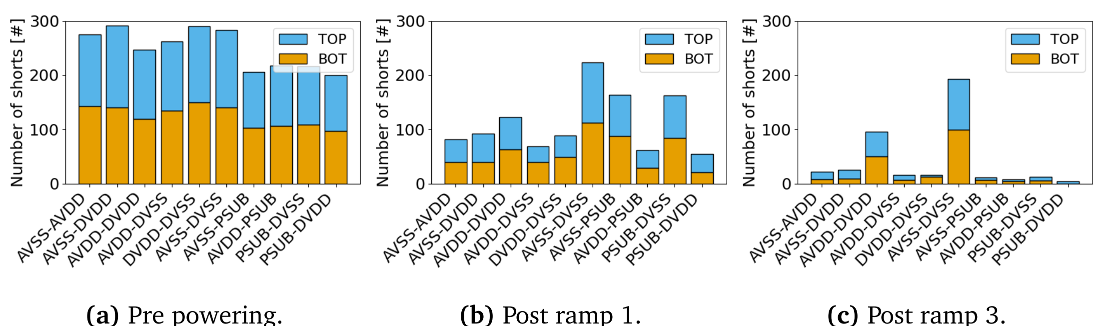
- AVSS–PSUB & PSUB–DVSS: The ground nets AVSS and DVSS, and the PSUB net are held at the same 0 V ground potential prior to ramp 3. Therefore, no burn-throughs will appear on AVSS–PSUB and PSUB–DVSS power net-pair combinations. HUs successfully pass the power ramp regardless of the presence of a short. The reduction in the number of observed shorts between pre-powering and post-ramp 1 is attributed to the reduction of parasitically observed shorts (cf. Section 6.2.5). After ramp 3, where PSUB is ramped to $-1.2\,\mathrm{V}$, AVSS–PSUB and PSUB–DVSS are burned through as well. The number of observed shorts each reduces to below 20 across all 1,600 measured HUs from the 80 MOSS sensors.

- AVSS–DVSS: The ground nets AVSS and DVSS of the analogue and digital domain, respectively, are always held at the same 0 V ground potential (off-chip). There are no burn-throughs observed on this power net pair. The reduction in observed AVSS–DVSS shorts is again due to a decrease in the number of parasitically observed shorts by opening real shorts with burn-throughs.



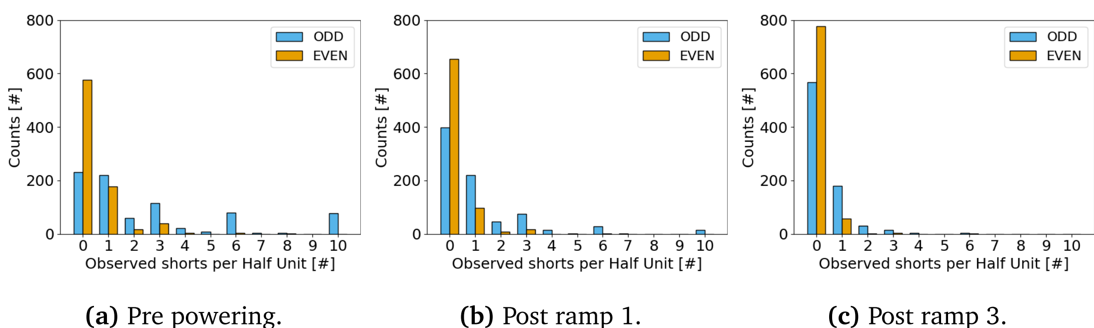
(a) Pre powering. **(b)** Post ramp 1. **(c)** Post ramp 3.

Figure 6.21: Observed shorts per wafer prior to powering (a), after ramp 1 (b), and after ramp 3 (c). The contribution of shorts in top and bottom HUs is equal, with counts quoted in the legends.



(a) Pre powering. **(b)** Post ramp 1. **(c)** Post ramp 3.

Figure 6.22: Observed shorts per affected power net-pair combination prior to powering (a), after ramp 1 (b), and after ramp 3 (c). Contributions of top and bottom HUs to the number of observed shorts for each power net pair are equal. See text for discussion of distinct patterns in (b) and (c).



(a) Pre powering. **(b)** Post ramp 1. **(c)** Post ramp 3.

Figure 6.23: Observed simultaneous shorts per HU before powering (a), after ramp 1 (b), and after ramp 3 (c). The fraction of no observed shorts increases, for both odd and even numbered wafers. The number of simultaneously observed shorts is strongly reduced, especially for odd-numbered wafers.

- AVDD–DVDD: Both AVDD and DVDD, the supply nets of the analogue and digital domains, respectively, are nominally powered to 1.2 V. During power ramp-up of a specific power net, the other domains are held at 0 V with the power supply.

> If a short exists between AVDD and DVDD, and, for example, the AVDD domain is powered up, the DVDD domain is pulled up as well, as the power supply does not provide a negative voltage to force the DVDD net to 0 V. Therefore, an AVDD–DVDD short can lead to parallel power-up of the analogue and digital domains. Shorts between AVDD and DVDD can, however, burn through if the current through the short is high enough to open it. After ramp 3, the number of remaining AVDD–DVDD shorts is approximately five times higher than for other power net-pair combinations (excluding AVSS–DVSS). An AVDD–DVDD short does not reduce the powering yield, however, as the two domains are operated at the same potential of 1.2 V.

The distribution of simultaneously observed shorts per HU is shown in Figure 6.23 before powering, after ramp 1, and after ramp 3, respectively. Here, distributions for odd and even numbered wafers are shown separately (as a deviating underlying failure density was established). The number of no shorts ('0') increases with the opening of short circuits. The counts of simultaneously observed shorts are strongly reduced, especially for odd-numbered wafers.

As stated above, after ramp 3, 84% of observed shorts are opened up. Excluding observed shorts on AVDD–DVDD and AVSS–DVSS power nets, which do not contribute to powering yield loss, 114 observed shorts remain after ramp 3 – a fraction of 5% of observed shorts prior to first powering.

In terms of HUs after ramp 3: 292/1600 HUs (18%) still exhibit at least one short. Excluding AVDD–DVDD and AVSS–DVSS shorts, this fraction reduces to 73/1600 (5%). The apparent discrepancy with the 13% of non-powerable units at PSUB $= -1.2\,\mathrm{V}$ (see Figure 6.16) is explained by the operational 25 mA current limit on the PSUB supply net, which may restrict successful powering without necessarily correlating with a remaining short[3]. In case of PSUB $= 0\,\mathrm{V}$, shorts on power net pairs AVSS-PSUB and PSUB–DVSS do not contribute to a powering yield loss, and a remaining

[3]Note: Technically, particularly on odd-numbered wafers, and from the observed failure density, there have to be shorts on some of the stitched M7 power nets spanning the full MOSS sensor. Shorts there are not directly measured and will likely cause increased leakage current, contributing to the yield loss.

61/1600 (4%) of HUs exhibit at least one short – precisely matching the overall powering yield loss in Figure 6.15.

6.5 Thermal camera results and short fault mechanism

By extracting the fault location using the thermal camera, it is possible to correlate the chip design locations with the shorted net pairs. As such, a detailed hypothesis on the short fault mechanism is formed – involving M7 and M8 metal layers – which is statistically tested and confirmed. Focused Ion Beam-Scanning Electron Microscopy (FIB-SEM) allows for (destructive) analysis of MOSS chips' metal stack cross sections. The hypothesised short faults at the extracted fault locations were found, confirming both the hypothesised failure mode and the failure analysis method developed in this work, which combines impedance measurements, power ramping, thermal camera, and chip design correlation for robust fault detection and classification. Given that burn-through currents are on the same order as operational currents, the present short-failure mode would have been missed without dedicated measurements, with the measurements presented providing powerful tools for future chip characterisation. A corresponding article can be found at [8], which was presented at [167]. The MOSS chip metal stack was manufactured as a custom configuration in a collaborative effort with the foundry, and the results of the failure analysis allowed for pinpointing the issue in a constructive dialogue.

An in-depth discussion of the thermal camera results, correlation with impedance and powering measurements, hypothesis formation and validation, and FIB-SEM cross-sectioning can be found in Chapter 7[4].

6.6 Yield extrapolation

The powering yield, i.e. fraction of HUs that are successfully powered, was previously discussed in Section 6.3.1. The short failure mode discovered is attributed to a processing fault and is not expected to impact future devices (see Chapter 7). Furthermore,

[4]This Chapter might not be visible to you, as access is (temporarily) restricted. Contact the author or Bodleian Libraries, Oxford, for more information.

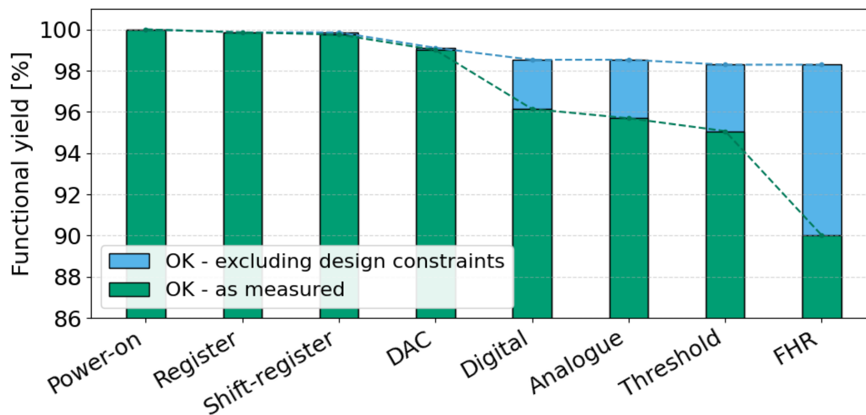


Figure 6.24: Yield drop-out plot for each step of the functional test sequence, excluding any powering yield loss due to shorts in the metal stack. The yield as measured, and corrected for known limitations of the design, is shown. Figure adapted from [7].

mitigation measures such as replacement and an adjusted routing strategy of the upper metal stack will additionally ensure the issue is eliminated. Hence, an extrapolation of the powering yield loss for the final ITS3 sensor (see also Section 3.6), based on the short-circuit failure mode, is not relevant.

It is, however, appropriate to discuss the functional yield, not impacted by short-circuit faults. An exhaustive measurement campaign using the functional test system, as outlined in Section 5.3, was conducted. All HUs classified as 'OK-I' and 'OK-II' after the thermal camera power ramp-up – termed 'powerable' – were tested (see Figure 6.16), and no correlation between burn-throughs ('OK-II') and any functional yield loss was observed. This is in full agreement with M7–M8 metal layer shorts, which, once burnt through, do not impact the operation of a given HU. The functional characterisation is detailed in [7]. A summary of yield losses per chip 'region' (cf. Figure 3.11) at each stage of the functional characterisation is given in Figure 6.24. The power-on stage here represents the power-on in the functional test system of only powerable HUs ('OK-I' and 'OK-II'), all of which pass this stage. The final functional yield, normalised to regions, is 90.0% and 98.3%, as measured, and excluding design constraints, respectively. The MOSS chip employs a simplistic readout architecture, resulting in pixel matrix yield loss during readout. These failures were expected at the design level, and are classified as design constraint contributions which will not

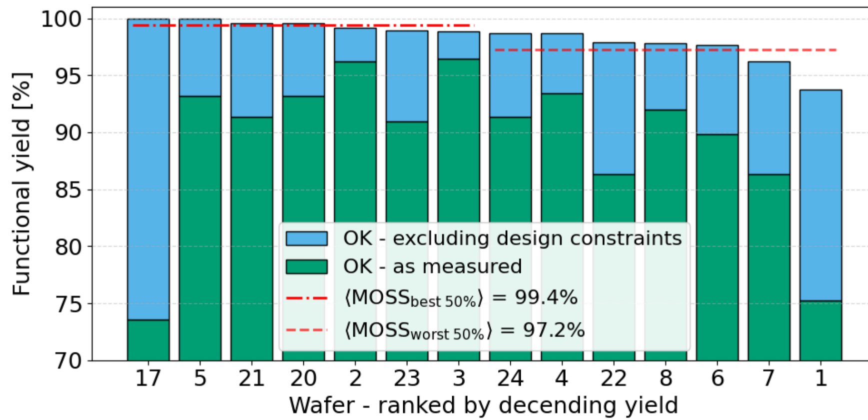


Figure 6.25: Per-region functional yield for each measured wafer, considering powerable HUs exclusively.

be featured in the final sensor design [7]. The best-case, average functional yield of all 80 tested MOSS sensors on the region level is therefore $Y_r = 98.3\%$. This is an optimistic view, as excluded regions due to design constraints could have other faults, and new failures in a future chip design are possible (however, as there are two more chip submissions, further design corrections can be implemented, and the MOSS functional yield excluding chip design constraints is used for extrapolation).

To illustrate wafer–wafer yield fluctuations, the ultimate functional region yield for each measured wafer is shown in Figure 6.25, and ranked from highest to lowest. The mean of the best half $\langle \text{MOSS}_{\text{best 50\%}} \rangle = 99.4\%$ and worst half $\langle \text{MOSS}_{\text{worst 50\%}} \rangle = 97.2\%$ is shown to indicate a range. The absolute number of powerable, and therefore functionally tested HUs and regions varies for each wafer (see powerable HUs, Figure 6.16).

From the MOSS functional yield, the required wafer lot size (accounting for the design granularity of the final chip) to build one ITS3 consisting of 6 layers with a certain dead fraction, is estimated:

- The MOSS region yield is, for an estimate, assumed to follow a Poisson distribution. The number of MOSS chips as a function of the fraction of regions with failures is roughly Poisson distributed (see Appendix B, including a correction to the Poisson model). No spatial gradient in the number of failures across wafer locations was observed. The area of the MOSS (excluding LEC and REC, which

were not tested here) is 10·(25.5 mm×14 mm), with 20·4=80 regions. The per-region area is therefore $A_r = 44.625\,\text{mm}^2$. The MOSS failure density D_M for a given region yield Y_r is then calculated as (following a basic Poisson-based yield model [110]):

$$D_M(Y_r) = -\frac{ln\,Y_r}{A_r}\,. \tag{6.9}$$

- The MOSS failure density is then scaled to the design granularity of the final ITS3 sensor with 144 tiles and a per-tile area of $A_t = 35.32\,\text{mm}^2$ (from a total chip area of 19.56 mm×260 mm excluding REC and LEC, cf. Section 3.6), giving a per-tile pass probability of

$$p_{tile} = e^{-D_M A_t}\,. \tag{6.10}$$

- The wafer-pass probability P_{wafer}, for a dead-fraction df, a total of 720 tiles per wafer, and the number of allowed dead tiles $k_{max} = \lfloor df \cdot 720 \rfloor$ is then:

$$P_{wafer} = \sum_{k=0}^{k_{max}} \binom{720}{k} (1 - p_{tile})^k \, p_{tile}^{720-k}\,. \tag{6.11}$$

  This represents an upper limit, as in the case of ITS3 layers L0 and L1, sections of the wafer can be selectively cut out and used – provided their local dead-fraction remains within acceptable limits – even if the overall wafer dead-fraction exceeds the specified threshold.

- Finally, the smallest wafer lot size of integer N, where $n_a = 6$ accepted wafers (with a dead fraction below df) are expected for a chosen confidence level C (e.g. 95%) is calculated as:

$$min(N):\ \sum_{n_a=6}^{N} \binom{N}{n_a} P_{wafer}^{n_a} (1 - P_{wafer})^{N-n_a} \ \geq\ C\,. \tag{6.12}$$

In Figure 6.26, the required wafer lot sizes for various dead fractions df of the final ITS3 sensor layers are shown, when extrapolating from the MOSS functional region yield. At the MOSS average scenario of a 98.3% region yield, which is considered as representative of transferable failures to the final sensor, 10 wafers suffice to build

one full ITS3 detector with 6 sensor layers at an ITS3 dead fraction of $\leq 2\%$. The target dead fraction for the final ITS3 is quoted as below 2% in the Technical Design Report; however, it is not considered a hard limit [6]. Given a sufficient wafer lot size, wafer-to-wafer fluctuations increase the probability of a subset of wafers with a higher yield. This is illustrated by the average worst-half $\langle \mathrm{MOSS}_{\mathrm{worst}\ 50\%} \rangle$ and best-half $\langle \mathrm{MOSS}_{\mathrm{best}\ 50\%} \rangle$ yield. A wafer lot size of $N = 20$ can therefore be considered a safe number to achieve a mean ITS3 dead fraction of less than 2%. Ultimately, it is foreseen to produce a wafer lot of $N = 50$, with the target of manufacturing two full ITS3 detectors (one spare). Considering that for layers L0 and L1 only parts of a full wafer are needed, and two more chip submissions are planned, the wafer lot size of $N = 50$ is a reasonable choice to achieve a $\leq 2\%$ ITS3 dead fraction.

The effect of dead tiles on the ITS3 performance in terms of pointing resolution and tracking efficiency is further discussed in Chapter 8, where higher dead fractions are additionally investigated.

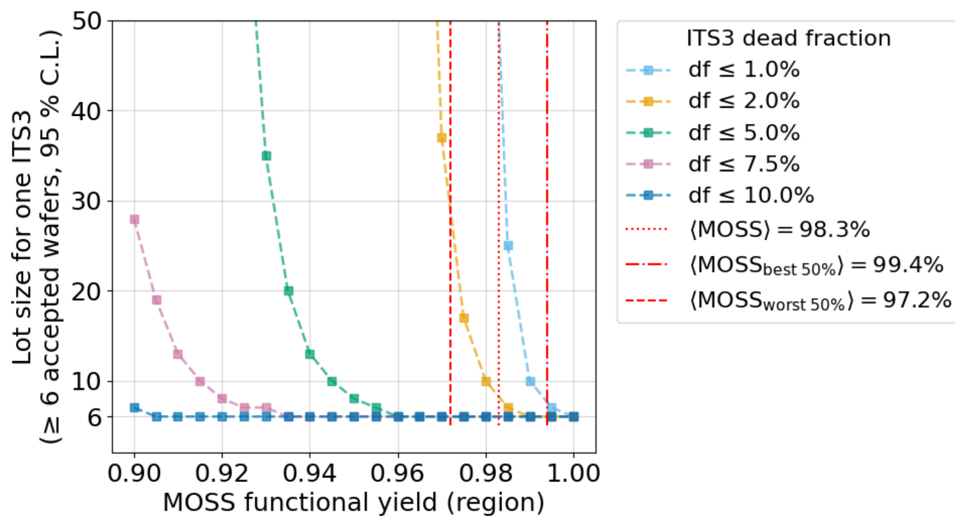


Figure 6.26: ITS3 yield extrapolation and required wafer lot size considering different accepted ITS3 dead tile fractions.

7
Failure and Root Cause Analysis

This chapter discusses in detail the correlation of hotspot locations, impedance measurements, and power ramps, forming the basis for a hypothesis regarding the failure mode. [redacted] This hypothesis is then tested statistically, validated, and examined in the context of the observed phenomena. The chapter introduces the Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) technique, which is used to investigate the failure mechanism further. To validate the hypothesis, a cross-section of the MOSS chip was prepared at the identified fault location, and the suspected defect was imaged using electron microscopy.

The chapter originally presented here cannot currently be made freely available via the Oxford University Research Archive (ORA).

Contact the author or Bodleian Libraries Oxford for further information.

8

Effect of Dead Areas on the ITS3 Physics Performance and Optimisation Strategy

This chapter discusses the effect of dead tiles in the final ITS3 sensors on the physics performance of the Inner Tracking System (ITS). After introducing the simulation framework, the impact on tracking efficiency and pointing resolution is discussed. A ranking method is introduced to assess the feasibility of using a Machine Learning (ML)-based model to predict the optimal sensor layer arrangement for the final detector. Finally, the impact of dead tiles on the Λ_c^+ reconstruction efficiency is discussed.

8.1 Simulation framework and procedure

The ITS3 detector performance simulation is based on the ALICE O^2 framework [72]. The following steps are performed to generate the data samples used in the performance study:

- The upgraded version of the ALICE ITS with the new ITS3 inner barrel is included when building the O^2 simulation framework. For the performance study, the `PYTHIA` 8.311 pp event generator [74] is used at $\sqrt{s} = 14$ TeV, and the beam pipe and ITS are included as detector components. In total, 400 batches of $n = 2000$ events were generated (with a 500 kHz interaction rate and 32 orbits at each

approximately 90 μs). The interaction point location varies along the z-axis with a Gaussian width of $\sigma_{IP} = 6$ cm. To assess the geometrical track acceptance and performance of the ITS3, the pp event generator was chosen over Pb-Pb to speed up the simulation process. It provides a pion $p_T - \eta$ spectrum covering the region of interest for the studies performed here.

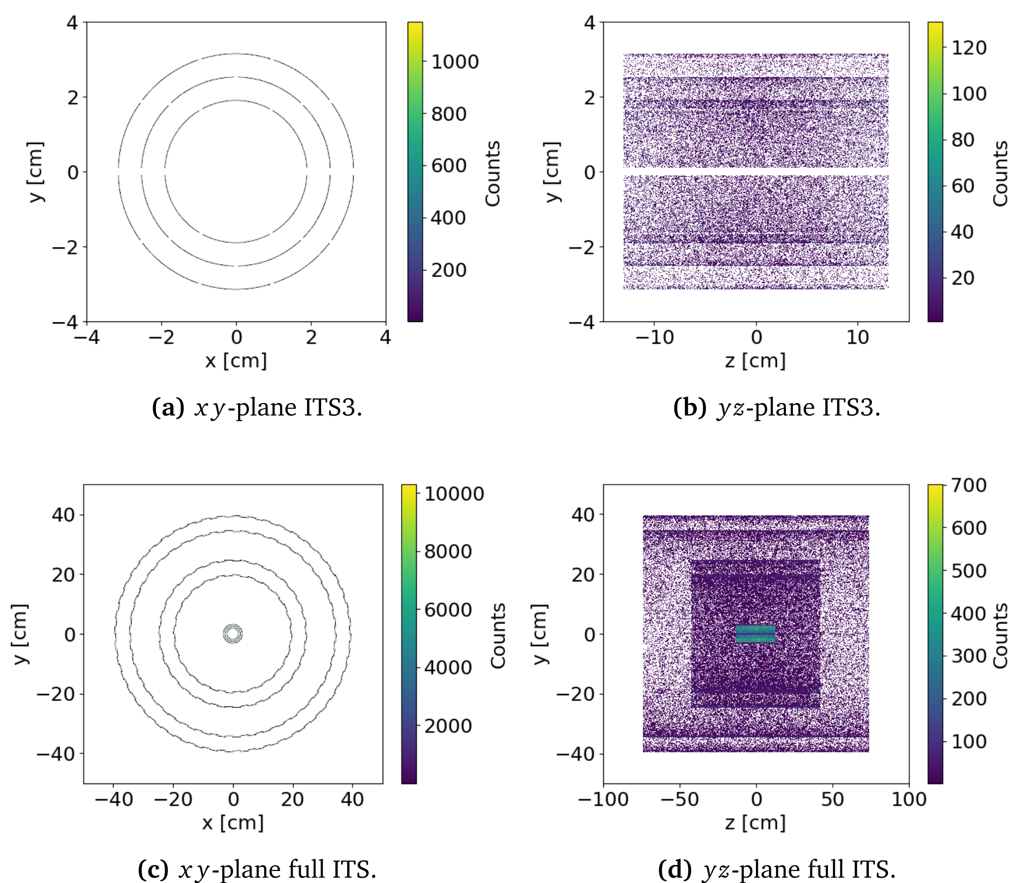
- The event digitisation is based on the pixel response of the currently installed pixel detectors, but mapped to the granularity and geometry of the ITS3. The full ITS3 geometry and material budget are taken into account, including physical gaps and non-sensitive on-chip areas. Cross-section views (xy-plane, yz-plane) of the digitised ITS3 inner barrel and full ITS barrel are shown in Figure 8.1. In this step, individual tiles of the ITS3 detector are optionally turned off, based on a so-called 'deadmap' (see below).

- Finally, clusterisation and tracking are performed in the reconstruction step. Longitudinal and transverse pointing resolution and tracking efficiency are calculated (see also Section 2.4.1).

8.1.1 Deadmap generation

To simulate the performance of the ITS3 in scenarios where random tiles are non-functional and switched off, multiple deadmaps were generated. Specifically, sets of 500 deadmaps were produced for each overall ITS3 dead fraction $df\,[\%] \in \{1, 2, 5, 10\}$. An example of one layer L1 sensor plane for the upper barrel half with a 2% dead fraction is shown in Figure 8.2.

The total number of dead tiles is constant for each dead fraction of 1, 2, 5, 10%, however, the distribution of dead tiles across the individual ITS3 detector layers (L0, L1, L2) naturally fluctuates between configurations.



(a) xy-plane ITS3.

(b) yz-plane ITS3.

(c) xy-plane full ITS.

(d) yz-plane full ITS.

Figure 8.1: Digitisation step: For one batch of $n = 2000$ events, the stored hits in the ITS3 and full ITS are shown. (a) and (b) illustrate the xy-plane inner barrel hits and full ITS barrel hits, respectively. The equatorial gap and gaps between sensor segments are visible in (a). (b) and (d) illustrate the recorded digitised yz-plane hits for the inner barrel and full barrel, respectively.

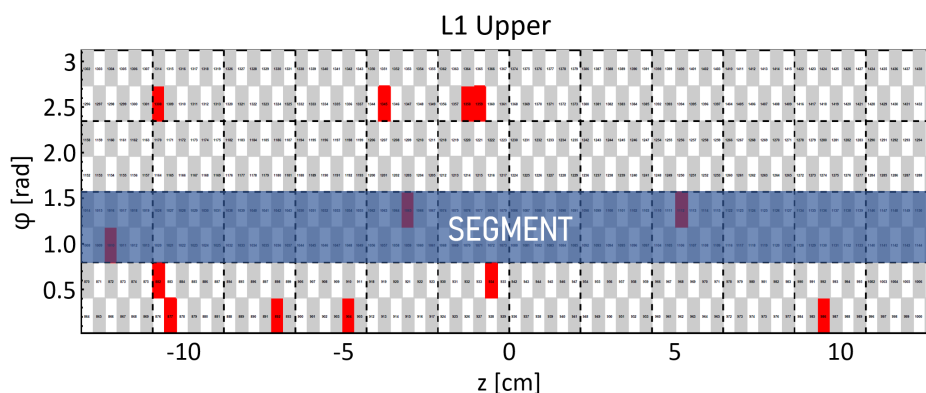


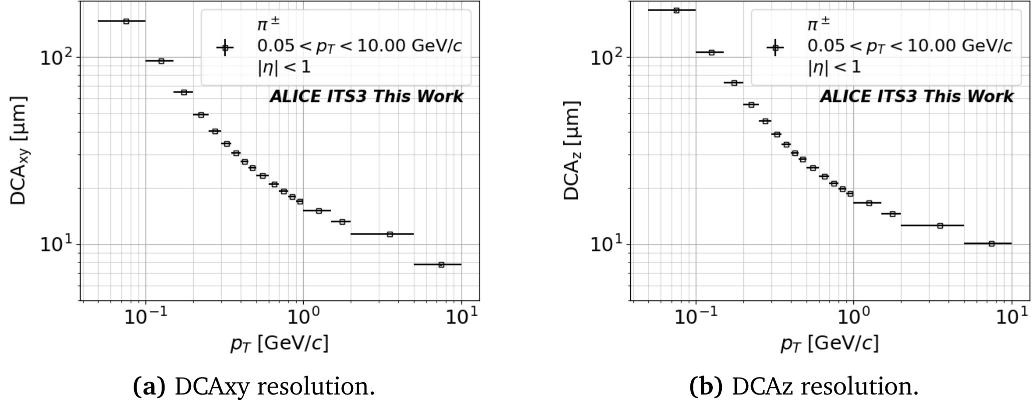
Figure 8.2: Example of a deadmap for the upper ITS3 L1 layer with an overall ITS3 dead fraction percentage of 2%. The dead tiles are marked in red. One segment, marked in blue for illustration, consists of 144 tiles. One L1 half-layer consists of 4 segments.

8.2 ITS3 pointing resolution and tracking efficiency

The simulated ITS3 pointing resolution and tracking efficiency are discussed for the case of fully functional sensor planes (i.e., with no deadmap applied). On-chip areas without pixel coverage are implemented (such as periphery, in between tile and RSU, and segment gap, cf. Section 3.6), as well as the mechanical equatorial gap (see Figure 8.1a). The following studies are based on primary charged pions, given their high abundance at transverse momenta $p_T < 10$ GeV/c, in line with current benchmarks [6]. Unless otherwise noted, a $|\eta| < 1$ cut is applied consistent with standard analyses.

8.2.1 Pointing resolution

The pointing resolution is computed for both the longitudinal z-plane (DCAz) and transverse xy-plane (DCAxy), analogous to Section 2.4.1. Combining 300 batches of simulations without deadmaps (entire ITS3 functional), the reference performance for $0.05 < p_T < 10.00$ GeV/c is extracted. Tracks with at least one hit in the ITS3 detector are selected to later ascertain the effect of dead tiles on the pointing resolution performance. If only tracks with three hits (clusters) within the ITS3 (hits in each layer) were selected, there would be no effect on the pointing resolution (albeit a lower number of overall tracks, which is discussed for the tracking efficiency degradation). The DCA of each track within pre-defined p_T slices is stored in two 2D histograms (DCA vs. p_T), separately for the longitudinal and transverse planes. For every p_T slice, the pointing resolution, now called $DCA_{\{xy,z\}}(p_T)$, is obtained as the width of a Gaussian fitted to the 1D projection of that slice. The corresponding bin error $\sigma_{DCA_{\{xy,z\}}}(p_T)$ is the 1σ uncertainty of the fitted width returned by the `Minuit` minimiser [169] within the used `ROOT` 6.32 framework [170]. The per p_T-bin DCAxy and DCAz resolution are shown in Figure 8.3a and Figure 8.3b, respectively.



(a) DCAxy resolution.

(b) DCAz resolution.

Figure 8.3: Pointing resolution in the transverse plane (a) and longitudinal plane (b) as a function of particle momentum. Pions are used as reference particles for momenta up to 10 GeV/c and $|\eta| < 1$.

8.2.2 Reconstruction efficiency

The achievable track reconstruction efficiency depends on the tracking algorithm. For the discussion of efficiency loss due to dead areas in the sensor, the 'Primaries Good' definition is used throughout (with pions as reference particles):

$$eff_{PrimariesGood}(p_T) = \frac{N_{reco,primary}(p_T)}{N_{generated,primary}(p_T)}\,, \tag{8.1}$$

where the denominator histogram represents every Monte Carlo (MC) primary pion within $|\eta| < 1$, and the numerator histogram contains all successfully ('good') reconstructed tracks. Fakes, which could inflate the efficiency, are ignored. Therefore, all clusters of the track belong to the same particle that produced them. Overall, four consecutive hits in the entire ITS are required for successful reconstruction. The resulting reconstruction efficiency is shown in Figure 8.4, with a fully functional ITS3 inner barrel. The binomial standard error per p_T bin is obtained, treating the distribution of reconstructed/not-reconstructed tracks as a binomial process.

8.3 Effect of dead tiles on pointing resolution and tracking efficiency

Using 100 batches of base simulations to speed up analysis, the digitisation and reconstruction steps were rerun, now including deadmaps. For each of the four

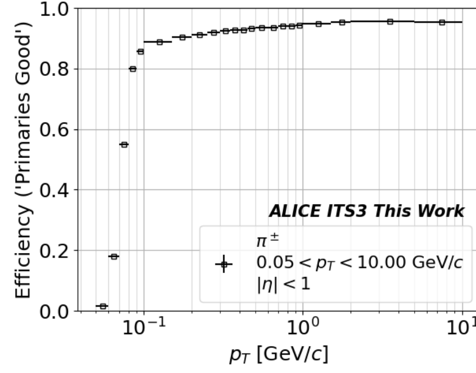


Figure 8.4: Tracking efficiency for the 'primaries good' criterion (see text). Pions are used as reference particles for momenta up to 10 GeV/c and $|\eta| < 1$.

investigated overall ITS3 dead fractions $df\,[\%] \in \{1, 2, 5, 10\}$, 500 random dead tile distributions (i.e., deadmaps) were generated. For each deadmap, the 100 base simulations were used to calculate the corresponding pointing resolution and reconstruction efficiency as introduced above. The resulting loss in performance is then obtained by comparison to the baseline (fully functional ITS3).

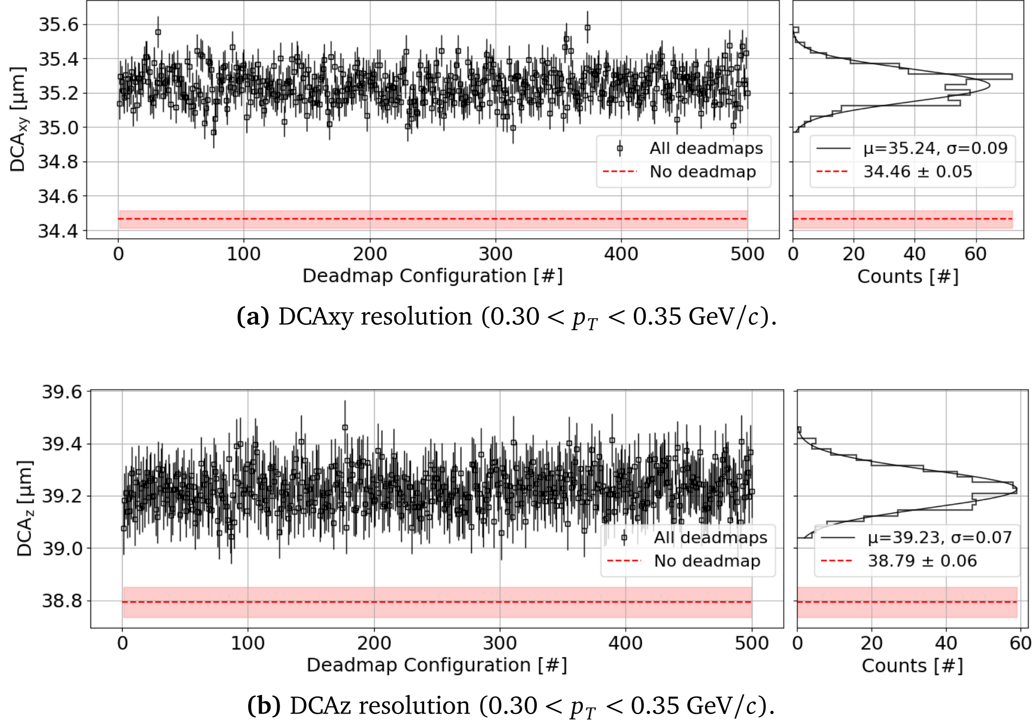
8.3.1 Pointing resolution degradation

For each deadmap configuration and each p_T-bin, the transverse and longitudinal pointing resolution are calculated (based on the 100 base simulations). The DCAxy and DCAz values for 500 deadmap configurations at a dead fraction of $df = 2\%$ are shown for $0.30 < p_T < 0.35$ GeV/c in Figure 8.5a and Figure 8.5b, respectively.

The DCAxy and DCAz values with deadmaps, $DCA^{deadmap}_{\{xy,z\}}$, follow a Gaussian distribution. The mean and standard deviation are extracted for each p_T-bin (same binning as in Figure 8.3) from a Gaussian fit. For each p_T-bin the pointing resolution loss is then calculated as a relative difference

$$L^{Rel.Diff.}_{\{xy,z\}}(p_T) = \frac{DCA^{deadmap}_{\{xy,z\}}(p_T) - DCA^{full}_{\{xy,z\}}(p_T)}{DCA^{full}_{\{xy,z\}}(p_T)} \,. \tag{8.2}$$

with $DCA^{full}_{\{xy,z\}}(p_T)$ the detector performance without dead tiles (no deadmap applied). The corresponding per p_T-bin uncertainty is calculated as



(a) DCAxy resolution ($0.30 < p_T < 0.35$ GeV/c).

(b) DCAz resolution ($0.30 < p_T < 0.35$ GeV/c).

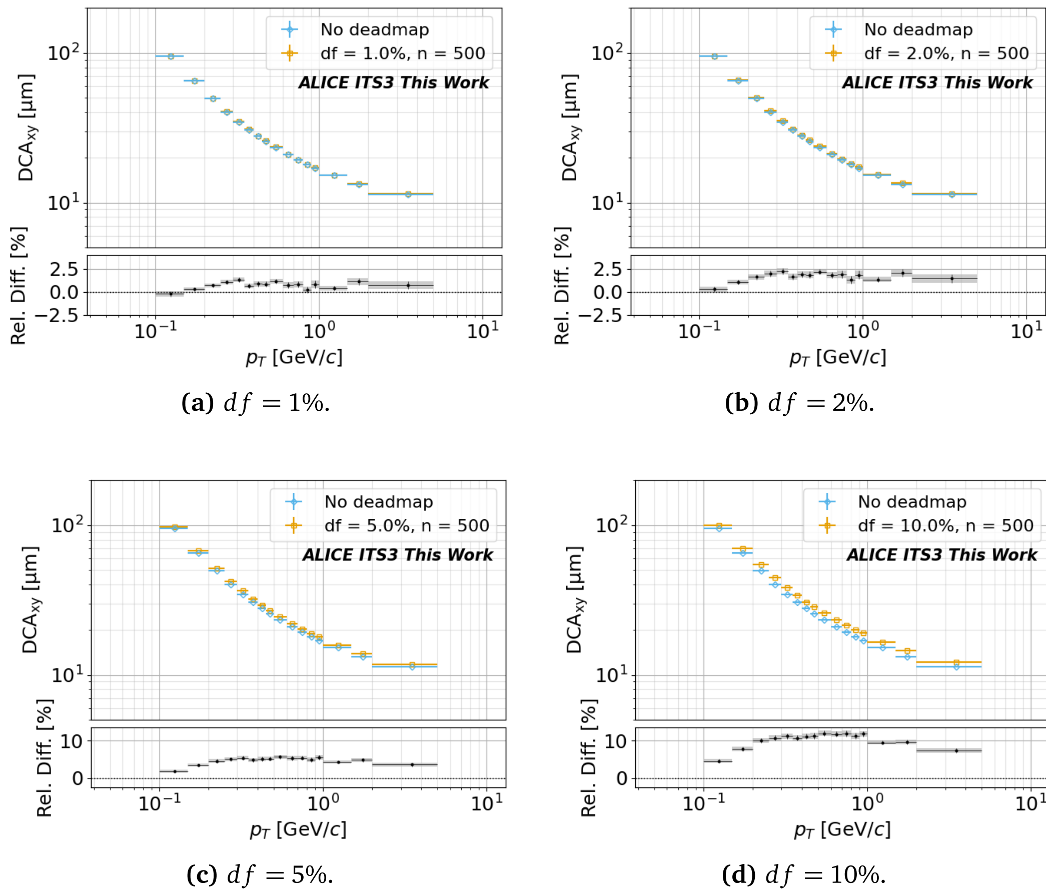
Figure 8.5: Pointing resolution in the transverse plane (a) and longitudinal plane (b) for 500 deadmap configurations at a dead fraction of 2%. A Gaussian fit is performed for all 500 configurations and shown on the right-hand side. The baseline performance without deadmaps (cf. Figure 8.3) is shown as a red band for the corresponding $0.30 < p_T < 0.35$ GeV/c p_T-bin resolution and uncertainty.

$$\sigma^2_{L_{\{xy,z\}}} = \left(\frac{\partial L^{Rel.Diff.}_{\{xy,z\}}}{\partial DCA^{deadmap}_{\{xy,z\}}} \sigma_{DCA^{deadmap}_{\{xy,z\}}} \right)^2 + \left(\frac{\partial L^{Rel.Diff.}_{\{xy,z\}}}{\partial DCA^{full}_{\{xy,z\}}} \sigma_{DCA^{full}_{\{xy,z\}}} \right)^2 \tag{8.3}$$

The resulting loss in pointing resolution in the transverse plane (DCAxy) for the studied dead fractions $df\,[\%] \in \{1,2,5,10\}$ is shown in Figures 8.6a, b, c, d, respectively. The loss in pointing resolution in the longitudinal plane (DCAz) is given in Appendix A.8.

At very low transverse momenta ($p_T \lesssim 0.2$ GeV/c), the pointing resolution is dominated by multiple scattering, and the impact of dead tiles on the achievable resolution is therefore reduced. The lowest ($p_T = 0.075 \pm 0.025$ GeV/c) and highest ($p_T = 7.5 \pm 2.5$ GeV/c) bins are statistically limited, given the smaller number of reconstructed tracks, and excluded from the analysis. In the range of $p_T = 0.2$–2.0 GeV/c, the loss in pointing resolution is directly proportional to the dead fraction

of the ITS3 – the dead fraction is about a 1:1 predictor of the worst-case pointing resolution degradation in the transverse plane at momenta below $O(5\,\text{GeV/c})$ and dead fractions of 1 to 10%. The loss in pointing resolution in the longitudinal plane is approximately 50% smaller: This can be attributed to the finer granularity of tiles in z-direction (260 mm/72 tiles $\simeq$ 3.6 mm per tile) compared to the larger arc length per tile in the transverse xy-plane (e.g. for L0: 120 mm (circumference)/12 tiles $\simeq$ 10 mm/tile). Furthermore, the solenoidal magnetic field bends charged particle tracks primarily in the xy-plane, while a straight-line fit suffices in the z-direction.



(a) $df = 1\%$.

(b) $df = 2\%$.

(c) $df = 5\%$.

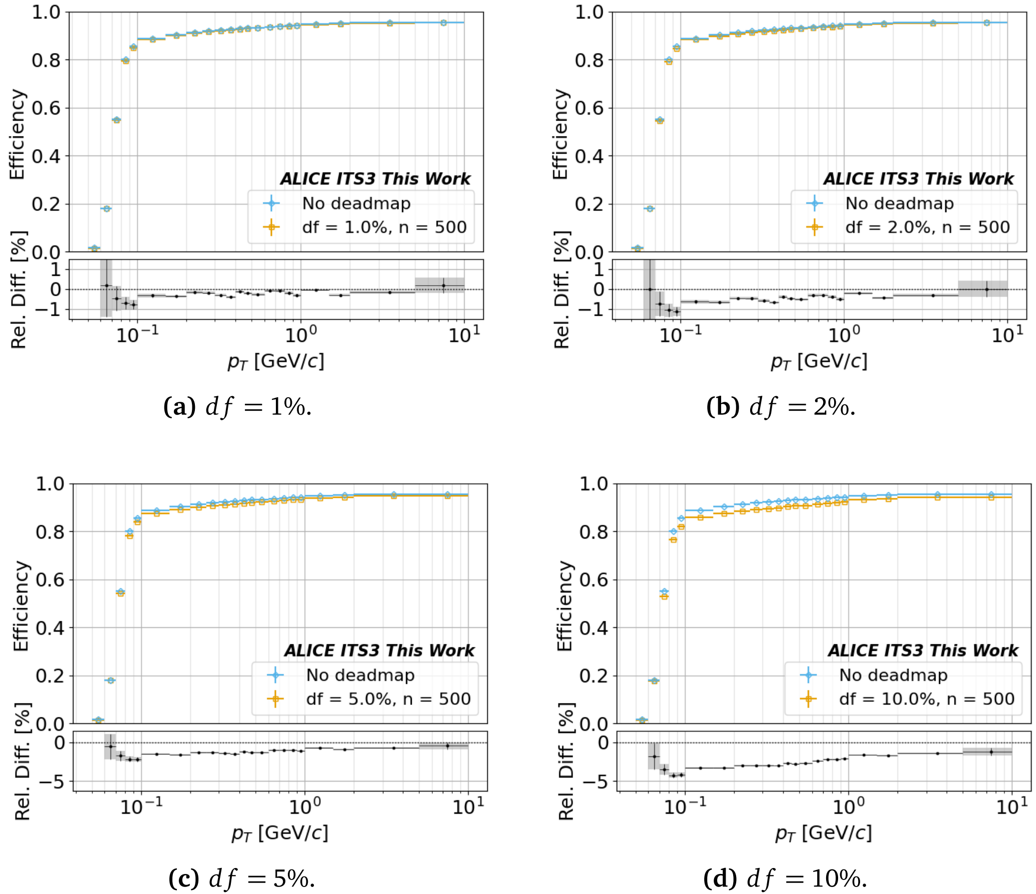
(d) $df = 10\%$.

Figure 8.6: Pointing resolution loss in the transverse plane for dead fractions of 1% (a), 2% (b), 5% (c), 10% (d). Each 500 random configurations were generated, with every configuration applied to 100 batches of base simulations during the digitisation step.

8.3.2 Reconstruction efficiency loss

In equivalent fashion to the pointing resolution study, the reconstruction efficiency loss is calculated for each simulated deadmap configuration and p_T-bin. A Gaussian fit is performed, and the resulting loss is shown in Figure 8.7. The reconstruction performance depends on the algorithm applied, which in the current implementation requires at least four consecutive hits in the full ITS. In the shown p_T range, tracks are lost if they are not successfully extrapolated to the vertex, and where hit information from the ITS3 is required. This effect is especially noticeable for low momenta $\lesssim 1\,\text{GeV}/c$.

Overall, the loss in reconstruction efficiency is observed to be less than 50% of



(a) $df = 1\%$. **(b)** $df = 2\%$. **(c)** $df = 5\%$. **(d)** $df = 10\%$.

Figure 8.7: Efficiency loss ('Primaries Good') for dead fractions of 1% (a), 2% (b), 5% (c), 10% (d). Each 500 random configurations were generated, with every configuration applied to 100 batches of base simulations during the digitisation step. Note that for (c) and (d), the lowest bin residuals lie below the limit of the plot axes.

the corresponding ITS3 dead fraction. For example, at the highest simulated ITS3 dead fraction of 10%, the observed loss in reconstruction efficiency is observed to range from approximately 2% to 4% for $0.1 < p_T < 10.0$ GeV/c. Momenta below $p_T < 0.1$ GeV/c are shown to illustrate the tracking efficiency turn-on, with the lowest bins suffering from a low number of entries, and carrying no significant information regarding reconstruction efficiency loss.

8.3.3 Deadmap ranking

A ranking of the deadmap configurations is studied in terms of transverse (DCAxy) and longitudinal (DCAz) pointing resolution, and reconstruction efficiency. The fractions of dead tiles in each of the ITS3 layers, and the percentage of dead tiles in each layer that lie in the outer region of the ITS3 detector, are studied. The outer region is here defined as the large-z part with 50% of the total layer area distributed within the left and right ends of the ITS3 detector, as illustrated for the upper L1 sensor plane in Figure 8.8.

Deadmaps are generated with well-defined overall ITS3 dead fractions (see Section 8.1.1), but varying fractions of dead tiles within each layer. Shown here are rankings for overall ITS3 dead fractions of 2% and 5%, and selected p_T bins.

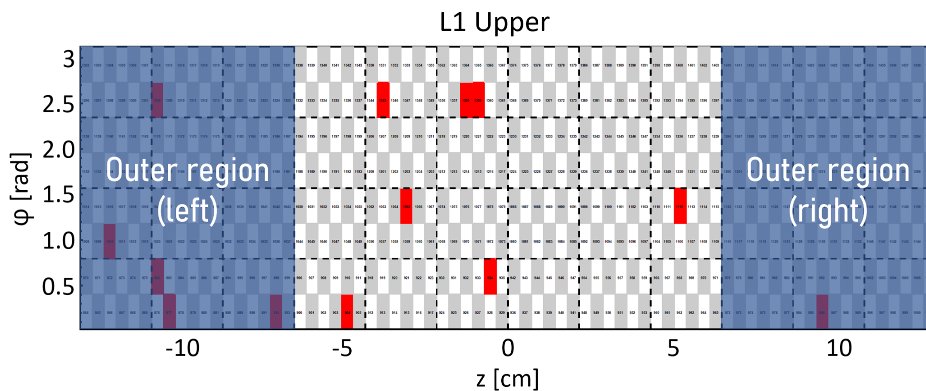
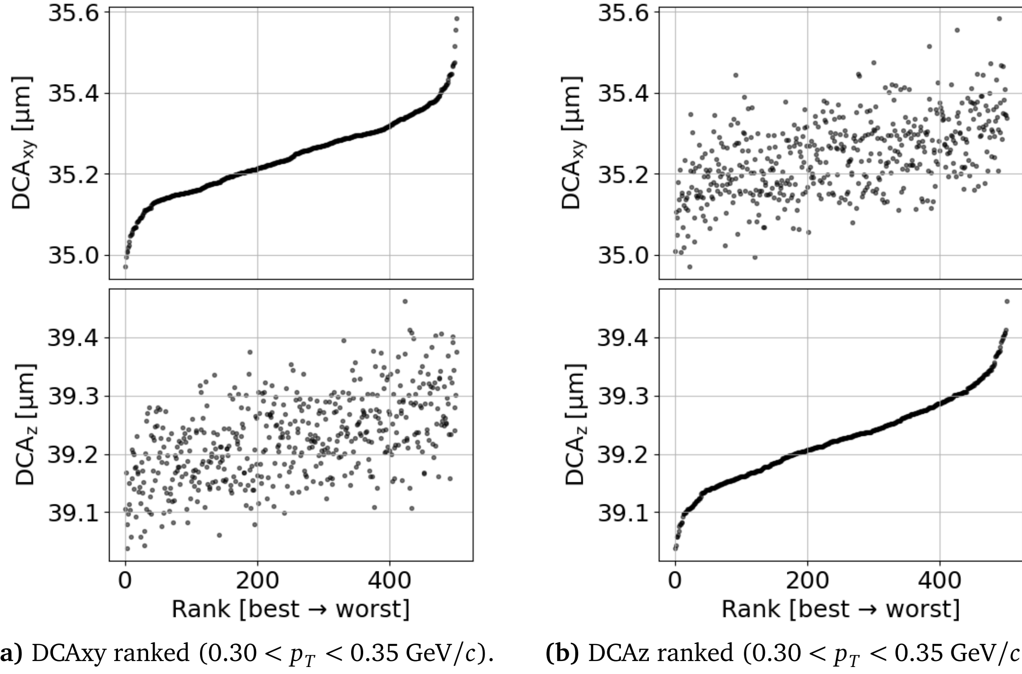


Figure 8.8: ITS3 outer region definition as used for each layer. The area within the outer region (left and right part) is 50% of the total layer area.

Ranking by pointing resolution

Each of the 500 deadmap configurations is ranked from best DCAxy or DCAz resolution to worst resolution for a given p_T-bin. Given the Gaussian distribution of

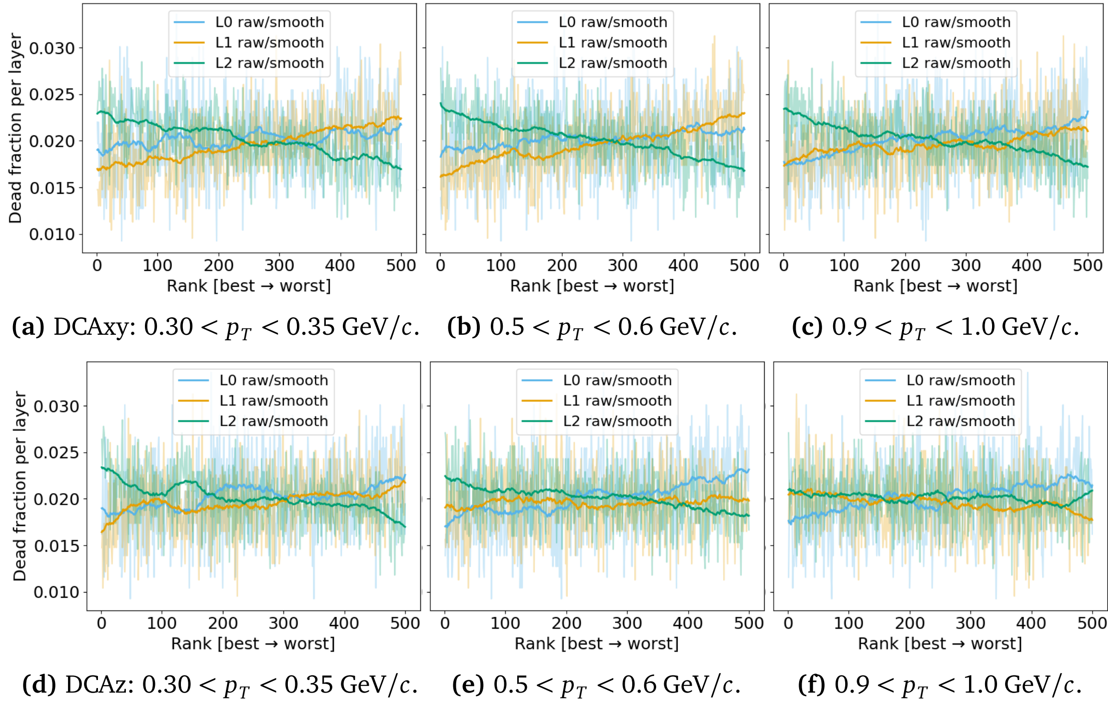


(a) DCAxy ranked ($0.30 < p_T < 0.35$ GeV/c). **(b)** DCAz ranked ($0.30 < p_T < 0.35$ GeV/c).

Figure 8.9: Ranking of 500 deadmaps for an overall ITS3 dead fraction of $df = 2\%$ each: Ranked by ascending (a) DCAxy, and (b) DCAz performance.

configurations (see Figure 8.5), a sigmoidal ranking is expected. As example, the deadmap rankings for a dead fraction of 2% and $0.30 < p_T < 0.35$ GeV/c are shown for both transverse (DCAxy) and longitudinal (DCAz) pointing resolution in Figure 8.9a and Figure 8.9b, respectively. For the DCAxy ranking, the corresponding DCAz resolution is shown, and vice-versa. A similar trend is observed, indicating a correlation between the two performance metrics.

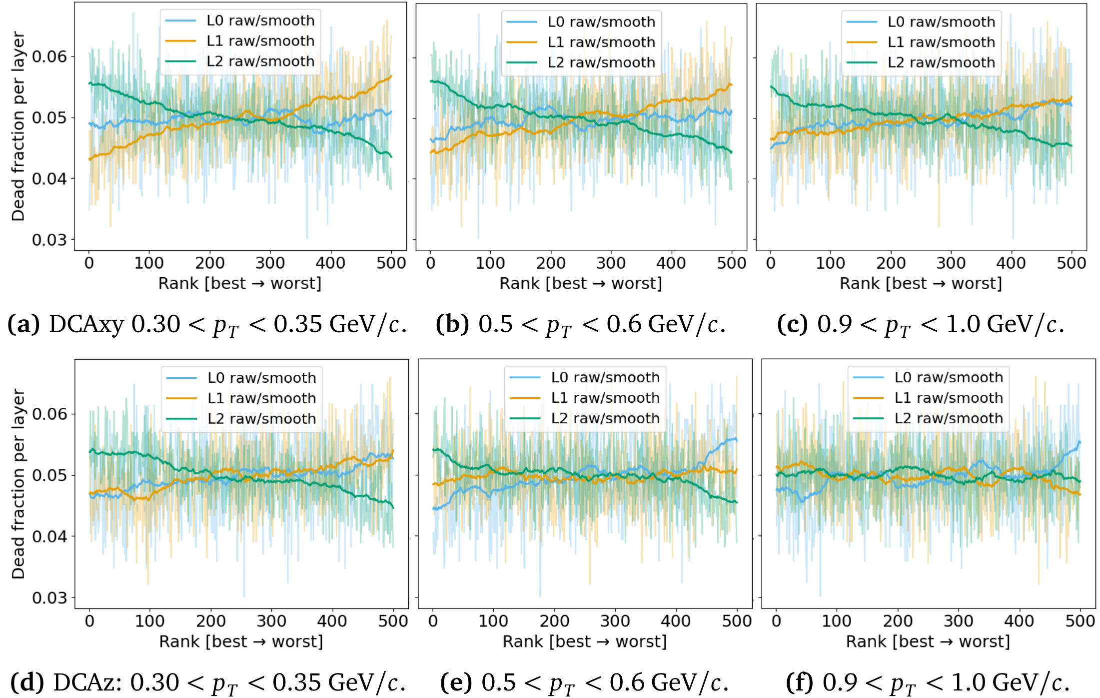
The fraction of dead tiles in each of the ITS3 layers (L0, L1, L2) when ranked according to the transverse and longitudinal pointing resolution for p_T bins $0.30 < p_T < 0.35$ GeV/c, $0.5 < p_T < 0.6$ GeV/c, $0.9 < p_T < 1.0$ GeV/c are shown in Figure 8.10 for an overall simulated ITS3 dead fraction of 2%. Similarly, in Figure 8.11, data is shown for an overall simulated ITS3 dead fraction of 5%. For better illustration of the trend, smoothed lines are overlaid on top of the single configuration dead fraction layer split. These are calculated as rolling-window means (window length 15) for each layer individually. As expected, the best-performing configurations show the lowest percentage of per-layer dead fractions in layers L0 and L1, with the highest layer dead fraction in L2 (the overall ITS3 dead fraction is constant across all configurations).



(a) DCAxy: $0.30 < p_T < 0.35$ GeV/c. **(b)** $0.5 < p_T < 0.6$ GeV/c. **(c)** $0.9 < p_T < 1.0$ GeV/c.

(d) DCAz: $0.30 < p_T < 0.35$ GeV/c. **(e)** $0.5 < p_T < 0.6$ GeV/c. **(f)** $0.9 < p_T < 1.0$ GeV/c.

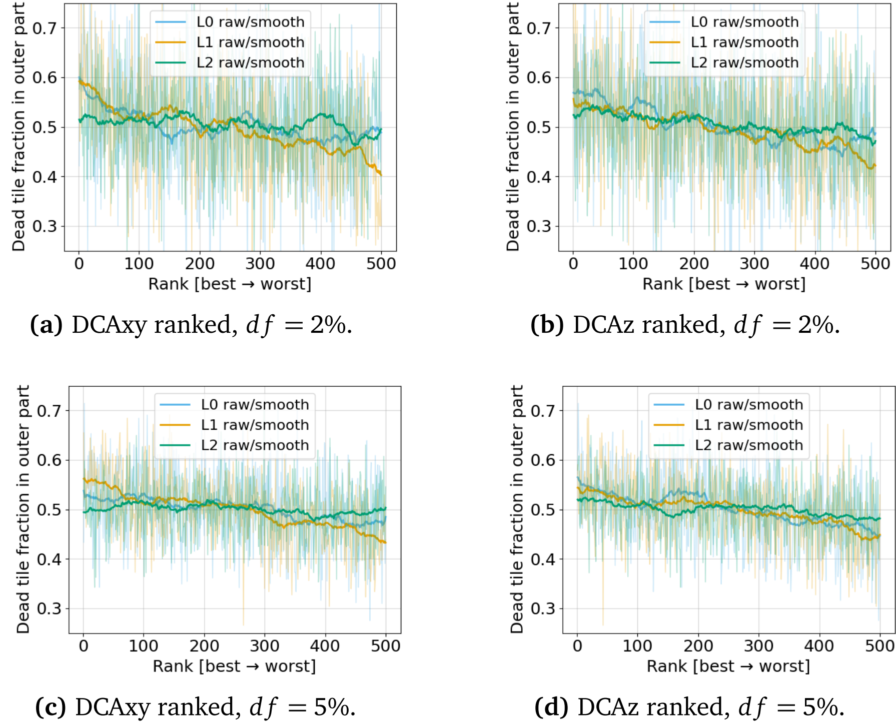
Figure 8.10: Per-layer dead fraction for an overall simulated ITS3 dead fraction of 2%, ranked according to DCAxy (a, b, c) and DCAz (d, e, f). The smoothed line represents a 15-unit rolling window.



(a) DCAxy $0.30 < p_T < 0.35$ GeV/c. **(b)** $0.5 < p_T < 0.6$ GeV/c. **(c)** $0.9 < p_T < 1.0$ GeV/c.

(d) DCAz: $0.30 < p_T < 0.35$ GeV/c. **(e)** $0.5 < p_T < 0.6$ GeV/c. **(f)** $0.9 < p_T < 1.0$ GeV/c.

Figure 8.11: Per-layer dead fraction for an overall simulated ITS3 dead fraction of 5%, ranked according to DCAxy (a, b, c) and DCAz (d, e, f). The smoothed line represents a 15-unit rolling window.



(a) DCAxy ranked, $df = 2\%$.

(b) DCAz ranked, $df = 2\%$.

(c) DCAxy ranked, $df = 5\%$.

(d) DCAz ranked, $df = 5\%$.

Figure 8.12: Fraction of dead tiles within each ITS3 layer located in the outer part. All data are at a momentum of $0.30 < p_T < 0.35$ GeV/c.

The trend is inverted for the worst configurations, where the least per-layer dead fraction is in L2. This trend is observed across the entire studied p_T-range, and all simulated ITS3 dead fractions, in agreement with the principles outlined in Section 2.3.

The same DCAxy and DCAz ranking sequences are used to analyse the distribution of dead tiles within the outer fraction of the ITS3 as defined in Figure 8.8. Four plots show a similar trend in Figure 8.12 at $0.30 < p_T < 0.35$ GeV/c. The results indicate that the overall pointing resolution is less affected when dead tiles are located in the outer regions of the ITS3. This is due to the lower number of reconstructed tracks at higher pseudorapidity $|\eta| > 0.8$, and the Gaussian distribution of reconstructed vertices around $z = 0$ with $\sigma = 6$ cm. With a lower probability of tracks originating from an interaction point far from $z = 0$, both ITS3 ends contribute less to the overall performance.

Additionally, the pointing resolution degrades at higher $|\eta|$, as tilted tracks traverse more material (higher MS contribution), and produce larger, more elongated hit clusters. The dead tile fraction in the outer section naturally shows higher fluctuations

for the overall ITS3 dead fraction of $df = 2\%$ compared to $df = 5\%$, given the smaller total dead tile count and therefore higher probability of observing a large fraction within or outside the ITS3 outer region.

From these observations, two qualitative conclusions can be drawn: to maximise ITS3 performance in terms of transverse and longitudinal impact parameter resolution, a detector configuration should be chosen where most dead tiles are (1) located in layer L2, and (2) within the outer regions of the detector along the z-direction.

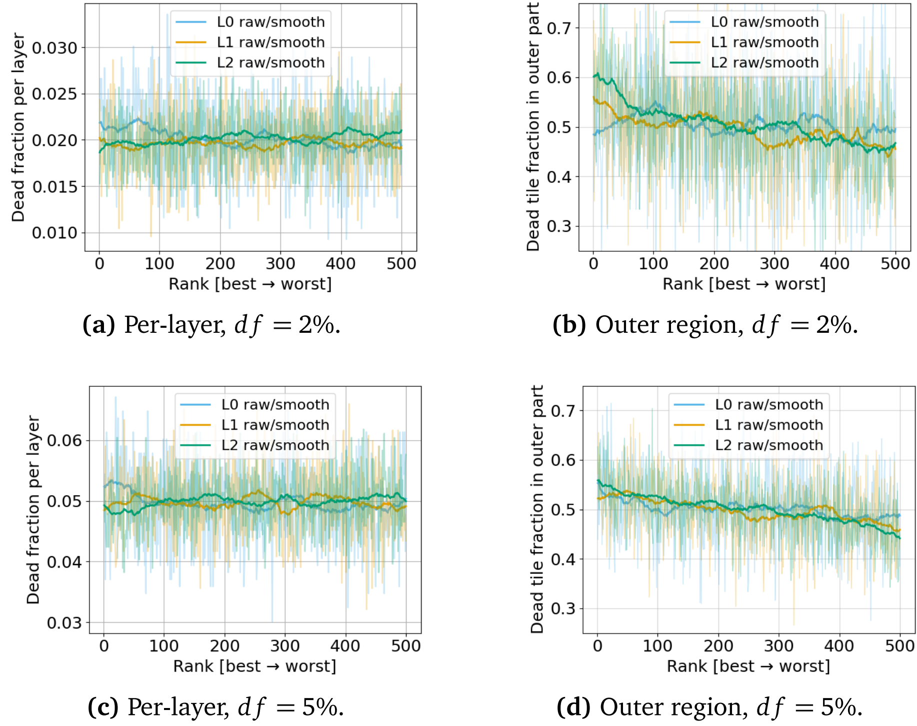
Ranking by reconstruction efficiency

Analogous to the ranking according to DCAxy and DCAz, a deadmap configuration ranking is performed based on reconstruction efficiency. The dead fraction per layer and dead fraction in the outer fraction are shown in Figure 8.13 for an overall ITS3 dead fraction of $df = 2\%$ and $df = 5\%$ at $0.30 < p_T < 0.35$ GeV/c. The per-layer dead fraction exhibits a flat trend, indicating that the ranked efficiency is independent of the number of dead tiles per layer. The track reconstruction efficiency is strongly algorithm dependent, with the layer in which a hit is recorded in the ITS3 contributing little to the overall performance. For the outer-region dead tile fraction, a similar trend as for the pointing resolution ranking is observed: the outer fraction of the ITS3 contributes less to the overall performance, due to the smaller number of tracks traversing those regions. Therefore, the best-ranked deadmap configurations show a higher number of dead tiles in the outer region for each layer.

8.3.4 Qualitative two-condition approach for detector optimisation

Two practical guidelines can be derived for allocating sensor planes with a given dead tile distribution to the ITS3 detector layers:

- As discussed in Section 2.3, to achieve the best impact parameter resolution, the innermost layer needs to be as close to the interaction point as possible (while minimising the material budget, reducing the MS contribution). Therefore, the sensor layers with the least dead fraction should be chosen for the innermost



(a) Per-layer, $df = 2\%$.

(b) Outer region, $df = 2\%$.

(c) Per-layer, $df = 5\%$.

(d) Outer region, $df = 5\%$.

Figure 8.13: Efficiency ranked data are at a momentum of $0.30 < p_T < 0.35$ GeV/c for a ITS3 dead fraction of 2% (a, b), and 5% (c, d). Per-layer dead fractions and outer region fractions are shown, respectively.

layer L0 for the best impact parameter resolution. This is in agreement with the trends visible for the $DCA_{\{xy,z\}}$ loss ranking, where the best configurations show the least dead tiles per layer in the innermost layers L0 and L1.

- For most physics analyses, a pseudorapidity cut at $|\eta| < 1$ is performed. The probability of tracks originating from an interaction point far from $z = 0$ decreases ($\sigma = 6$ cm), such that both ends (in z-direction) of the ITS3 detector contribute less to the overall performance. Dead tiles in these regions, therefore, have a lesser impact on the overall ITS3 performance. This trend is observed both for $DCA_{\{xy,z\}}$ loss and reconstruction efficiency loss rankings of deadmap configurations.

In summary, a qualitative allocation strategy is to select wafers such that the innermost ITS3 layers contain the fewest dead tiles, particularly in the central ($z \approx 0$) region of the barrel.

8.4 Optimisation of ITS3 layer geometry using an artificial neural network

The selection of sensor layers for the final ITS3 detector poses an interesting optimisation problem. A manufacturing lot of $O(50)$ wafers will be produced for the final detector and tested with a wafer probing system. This process will identify which wafers, and parts of wafers, have the highest functional yield. The question then is: how should wafers be diced to create layers L0, L1, L2 for a full ITS3, and how should the layers be arranged to maximise the ITS3 performance? The combinatorial space of possible sensor layer arrangements is large, making this a non-trivial optimisation task.

To explore one possible route toward geometrical optimisation, an Artificial Neural Network (ANN) surrogate model of the detector is employed in the discussion that follows.

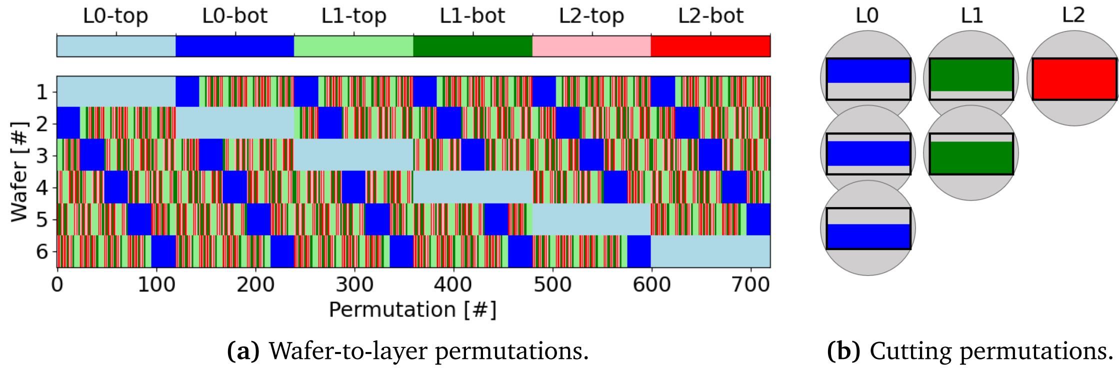
8.4.1 Number of permutations of ITS3 layer arrangements

Six wafers are needed to create one ITS3 barrel[1] (cf. Figure 2.8b): two each for layers L0, L1, and L2. The number of permutations for assigning a given wafer to a specific layer is $n_{wp} = 6! = 720$, as illustrated in Figure 8.14a.

Additionally, there are 2 and 3 possible ways to cut layer L1 and L0 out of a wafer, respectively. This is illustrated in Figure 8.14b. Since two sensors per layer are needed, the total number of cutting permutations is $n_{cp} = (3)^2_{\mathrm{L0}} \cdot (2)^2_{\mathrm{L1}} \cdot (1)^2_{\mathrm{L2}} = 36$.

For a single set of 6 wafers needed to manufacture one ITS3 detector, a total of $n_{tp} = n_{wp} \cdot n_{cp} = 25,920$ permutations of wafer-to-layer and layer cutting exist. This number is the number of deadmap configurations that would need to be simulated as shown above, to extract the best-performing configuration. Currently, on a machine with 48 cores and 128 GB RAM, 100 deadmap simulations (threaded on 48 cores) take about 10 hours to run. It is therefore feasible to simulate all permutations for a single set of 6 wafers (∼11 days, and less with increased compute power).

[1] Ultimately, two fully functional ITS3 barrels are planned to be produced (altogether 12 sensor planes); however, the present discussion focuses on a single ITS3 barrel, with the approach being scalable to both barrels.



(a) Wafer-to-layer permutations. **(b)** Cutting permutations.

Figure 8.14: (a) Graphical representation of all 720 permutations of wafer-to-layer allocation for 6 sensor layers made from 6 distinct wafers. Each column represents one permutation, where each wafer is allocated to a unique layer. E.g. for the first 120 permutations, wafer #1 is allocated to layer L0-top, and all other wafers are varied in layer allocation. (b) Wafer cutting permutations for L0 (3 ways), L1 (2 ways), and L2 (1 way) sensors.

However, the six wafers will be selected from a larger production lot, which results in a much larger combinatorial space.

For example, if 10 wafers meet the basic requirements to be considered as potential ITS3 layers, the total number of layer arrangement permutations becomes: $\binom{10}{6} \cdot n_{tp} = 5,443,200$. For 20 accepted wafers, this number increases dramatically to 1,004,659,200. It is therefore not feasible to run the full simulation process for every permutation.

After introducing the qualitative two-condition approach in Section 8.3.4, an alternative route is explored here: using a surrogate model to predict the layer geometry with optimal performance (for a given performance parameter, see below)

8.4.2 Neural network model training

For this feasibility study, one parameter – the loss in pointing resolution in the longitudinal plane DCAz – was taken as a measure to maximise the ITS3 performance. An artificial neural network, more specifically a Deep Neural Network (DNN), was trained on the simulated deadmaps and corresponding DCAz loss $L_{DCA_z} = (DCA_z^{deadmap} - DCA_z^{full})/DCA_z^{full}$ (this is not the 'loss function' of the DNN, but the predicted performance parameter). To simplify the input space, the deadmap was binned such that each group of 12 tiles (i.e., one RSU) was treated as a single unit.

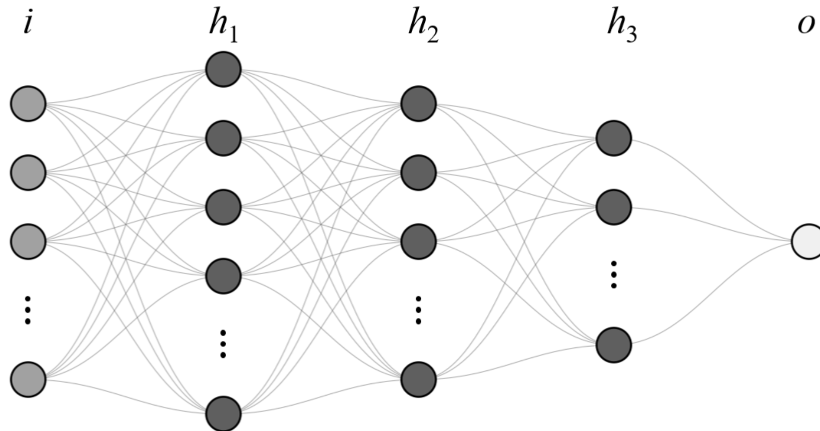


Figure 8.15: Schematic illustration of the DNN with input layer i, three hidden layers h_i, and scalar output o. The input layer is concatenated from three separate input vectors for each ITS3 layer (see text).

This reduces the input parameter space by a factor of 12 (in an extended version of the model, the full granularity could be conserved). The loss in impact parameter resolution in the range of 0.3 GeV/c to 0.5 GeV/c was chosen (where the loss is approximately flat), eliminating the need to include p_T as an additional input feature. Altogether 3000 fully simulated deadmaps with dead fractions of 2, 5, 10% are used to train the model[2]. The model has 288 input parameters (3456:12 binned ITS3 tiles), which are split into input vectors (branches) for each ITS3 layer (L0: 72, L1: 96, L2: 120). The model has one scalar output, the relative DCAz loss, on which it is trained.

The model architecture is a multi-branch multilayer perceptron (MLP) [171]. A conceptual diagram is shown in Figure 8.15. The three input branches (hence, 'multi-branch'), corresponding to the three ITS3 layers, allow for tuning the model capacity (and regularisation) for each layer individually (while allowing to probe activations of each layer-specific path). The outputs of the three branches are concatenated into a single combined input layer i with 208 features (after branch-specific pre-processing: $L_0 \rightarrow 80 + L_1 \rightarrow 64 + L_2 \rightarrow 64$ features). This vector is forwarded through a stack of three fully connected hidden layers h_i ('deep network') with decreasing sizes of

[2]This dataset was generated with an O^2 version containing a bug with a cut-off ITS3 volume in z-direction by about 3 cm, with negligible impact on this feasibility study

$256 \to 128 \to 64$ nodes. An artificial neural network with more than one or two hidden layers is commonly referred to as a deep neural network (DNN) [172, 173].

Each layer is followed by batch normalisation (BN), rectified linear unit activation function (ReLU), and a 10–20% dropout during training (see below) [174–176]. The final linear node o produces the learned residual correction $\Delta f_\theta(x)$ as one scalar output. The network predicts, for a given deadmap configuration x, the dimensionless relative DCAz loss:

$$\hat{y}(x) = f_\theta(x) = \underbrace{\Delta f_\theta(x)}_{\text{learned residual}} + \underbrace{E(x)}_{\text{empirical tuning}}, \quad (8.4)$$

where

$$E(x) = [e^{-8df} + 0.8] \sum_{j=0}^{2} w_j \, (df_L)_j \quad (8.5)$$

with overall ITS3 dead fraction df, layer dead fraction df_L and layer weight w. The empirical tuning term was introduced (similar to e.g. [177]) to improve the learning of deadmap variations for low overall dead tile fractions, resulting in an approximately 15% improvement in prediction accuracy during training.

The model is trained by minimising an uncertainty-weighted mean-squared error:

$$L_{\text{err}}(\theta; x_i, y_i, \sigma_i) = \frac{[f_\theta(x_i) - y_i]^2}{\sigma_i^2}, \quad (8.6)$$

where x_i is the deadmap configuration, y_i is the ground-truth relative DCAz loss, and σ_i the corresponding uncertainty. The goal of the DNN training is to find a set of learnable parameters $\theta^\star$, which minimises L_{err} on the training data, while generalising to unseen deadmap configurations. The learnable parameter set $\theta = \{W_1^i, b_1^i, \dots, W^{h_1}, b^{h_1}, \dots, W^o, b^o\}$, contains the weight matrices W^l and bias vectors b^l of the entire network, for inputs i (three branches 1–3), hidden layers h_1, h_2, h_3, and final output layer o. Weight matrices encode the connection strengths between consecutive layers l, with the corresponding bias vectors providing per-unit offsets to steer node activation. Together they implement the affine transformation for a hidden layer as $h_l = \phi\,(W^l h_{l-1} + b^l)$, with the ReLU activation function $\phi(z) = \max(0, z)$ (BN

is applied afterwards, and omitted here for clarity).

The DNN is implemented with the PyTorch framework [178]. The full available dataset $\mathscr{D}$, based on $N = 3000$ deadmap configurations and relative DCAz loss for $0.3 < p_T < 0.5$ GeV/c, is split into a final test set $\mathscr{F}$, a training set $\mathscr{T}$, and a validation set $\mathscr{V}$. The partitions are:

$$\mathscr{D} : \{(x_i, y_i, \sigma_i)\}_{i=1}^{N=3000} = \mathscr{F} \cup \mathscr{T} \cup \mathscr{V}, \ |\mathscr{F}| = 0.1N, \ |\mathscr{T}| = 0.72N, \ |\mathscr{V}| = 0.18N. \tag{8.7}$$

All partitions are randomly drawn. The final test set $\mathscr{F}$ is set aside to evaluate the fully trained model. The training set $\mathscr{T}$ and validation set $\mathscr{V}$ are used to train and validate the DNN. For every training step t ('epoch'), the training loss $\mathscr{L}_{train}^{(t)}$ and validation loss $\mathscr{L}_{val}^{(t)}$ are calculated as follows:

- Training loop: The entire training set $\mathscr{T}$ is shuffled and partitioned into K mini-batches $\{\mathscr{B}_t\}_{k=1}^{K}$ of size $m = |\mathscr{B}_t^{(k)}| = 64$. The per-batch loss is then:

$$L_{\text{batch}}(\theta; \mathscr{B}) = \frac{1}{m} \sum_{x_i, y_i, \sigma_i \in \mathscr{B}} L_{\text{err}}\left(\theta; x_i, y_i, \sigma_i\right). \tag{8.8}$$

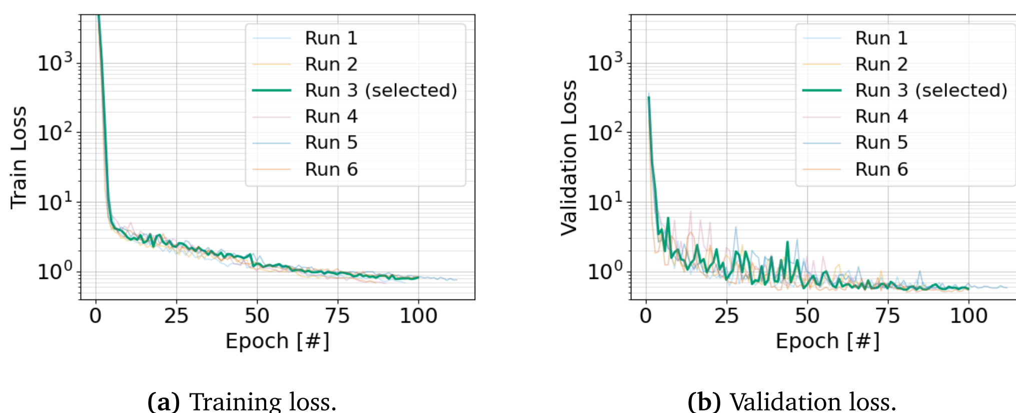
  Each mini-batch is processed once per epoch t, with $K = |\mathscr{T}|/m$ iterations per epoch. One iteration refers to one optimiser update on one mini-batch. The Adam optimiser is used with a default weight decay of $\lambda = 10^{-4}$ and starting learning rate $\eta = 5 \cdot 10^{-4}$ [179]. Each iteration, conceptually, performs the parameter update

$$\theta \leftarrow \theta - \eta(\nabla_\theta L_{\text{batch}}(\theta; \mathscr{B}) + \lambda\theta). \tag{8.9}$$

  The epoch-level training loss is the arithmetic mean of (8.8) over all K mini-batches of epoch t:

$$\mathscr{L}_{\text{train}}^{(t)} = \frac{1}{K} \sum_{k=1}^{K} L_{\text{batch}}\left(\theta_t; \mathscr{B}_t^{(k)}\right) \tag{8.10}$$

  Note: During the training loop, a 'dropout' is active, randomly masking 10-20% nodes in the DNN. Each mini-batch, therefore, sees a different thinned network, preventing any single sub-network from dominating, and ultimately leading to better generalisation.



(a) Training loss.

(b) Validation loss.

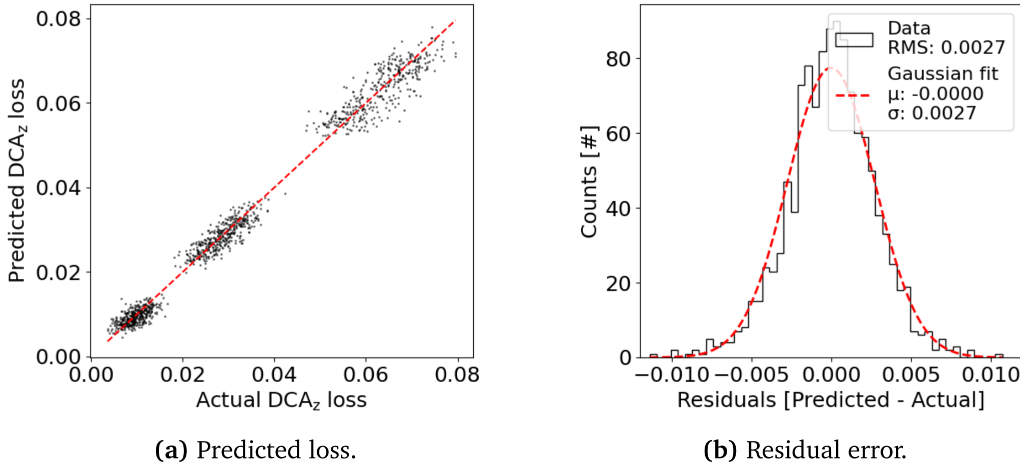
Figure 8.16: Examples of training loss (a) and validation loss (b) curves, with the selected model highlighted.

- Validation step: At the end of each epoch, the error is calculated on the entire validation set as ('dropout' is inactive)

$$\mathscr{L}_{\text{val}}^{(t)} = \frac{1}{|\mathscr{V}|} \sum_{(x_i, y_i, \sigma_i) \in \mathscr{V}} \frac{[f_{\theta_t}(x_i) - y_i]^2}{\sigma_i^2}. \tag{8.11}$$

  The validation loss is a deterministic estimate of the generalisation error. During training, it is monitored to both adjust the learning rate η if stagnation is detected, and stop the training loop when the validation loss plateaus.

Training loss and validation loss for each epoch of the model training are shown for a selection of six runs in Figure 8.16a and Figure 8.16b, respectively. The higher absolute training loss stems from the 10 to 20% dropout. The oscillations in the validation loss curve are expected, since it is computed once per epoch (rather than averaged over mini-batches), on a smaller dataset. The oscillations reduce after epoch 50, where the learning rate η is halved for the first time (after stagnating average improvement of the validation loss). For the six runs illustrated, the learning rate and the patience parameters of the scheduler and early stopping were varied as hyperparameters (see below). For each fully trained candidate model, the set-aside final test set $\mathscr{F}$ is used to evaluate the model performance. In Figure 8.17a, the performance of the selected model (highlighted in 8.16) is shown. For previously never-seen deadmaps, the model predicts the relative DCAz loss, which is compared to the actual DCAz loss (as simulated). As shown in Figure 8.17a, the predictions



(a) Predicted loss.

(b) Residual error.

Figure 8.17: (a) Predicted and actual relative DCAz loss for the set-aside final test-set $\mathscr{F}$. The red dashed line indicates perfect correlation. The three distinct data clusters correspond to a 2, 5, 10% ITS3 dead fraction, respectively. (b) Residual error on the prediction, with a Gauss fit overlaid. No bias ($|\mu| < 10^{-4}$) and a $\sigma = 0.27\%$ error spread are observed.

are distributed along the perfect-correlation line (dashed red), demonstrating strong agreement. Three clusters are visible, corresponding to an overall ITS dead fraction of 2%, 5%, and 10%, respectively. The residual error is shown in Figure 8.17b, following a Gaussian distribution centred at zero ($|\mu| < 10^{-4}$) with width $\sigma = 0.0027$. The model, therefore, does not introduce an upward or downward bias while successfully generalising to unseen data. The selection of the DNN model was based on the lowest bias and lowest error width during evaluation of the final test-set $\mathscr{F}$.

The model successfully demonstrates the prediction of the ITS3 performance from deadmaps in terms of relative DCAz loss within $0.3 < p_T < 0.5$ GeV/c.
This DNN model serves as a feasibility study, demonstrating the potential of a surrogate model approach. Dedicated hyperparameter optimisation, extended model validation, and possible model enhancements are the appropriate next steps to improve performance further.

Hyperparameters are settings, fixed before training a model, and include sizing of branch and layer widths, mini-batch size m, dropout rates, empirical tuning parameters, weight decay λ, (initial) learning rate η_0, patience parameters for learning rate adjustment and training stopping. For the discussed model, a convenient approach of a small grid-search was chosen, and six (out of $\sim$ 200) runs with optimisation

of learning rate and patience parameters are shown in Figure 8.16. Additionally, multiple extensions to the model can be considered, including the addition of the transverse pointing resolution DCAxy and tracking efficiency as output parameters. The training data set should be extended to include the full range of dead fractions between 1 to 10% in smaller intervals, with a larger variance between layers. The p_T dependency should be considered, and multiple particle species could be included, depending on the objective.

8.4.3 Deadmap ranking with the artificial deep neural network

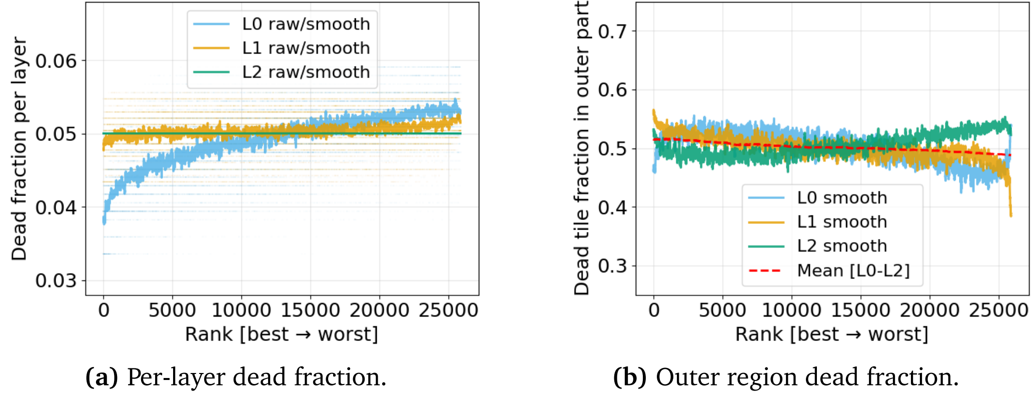
The DNN model performance is now further evaluated in terms of the practical use case of wafer-to-detector geometry allocation. Here, individually simulated wafers are the basis for deadmap configurations: Consider a set of 6 wafers, each with 720 tiles needed to construct a full ITS3 detector. For each wafer, a dead tile distribution is simulated once (for a per-wafer defined dead fraction df_W). From the set of per-wafer dead tiles, all possible wafer-to-layer and cutting permutations (as discussed in Section 8.4.1) are computed. Hence, for 6 wafers with fixed dead tile pattern, $n_{tp} = 25,920$ deadmaps are generated and evaluated.

Multiple scenarios are illustrated and discussed below. A constant per-wafer dead fraction is discussed (for 6 wafers), followed by a variable per-wafer dead fraction (for 6 wafers). To demonstrate the feasibility to evaluate a large combinatorial space, cases for 8 wafers (with $\binom{8}{6} \cdot 25,920(n_{tp}) = 725,760$ deadmaps) and 10 wafers (with $\binom{10}{6} \cdot 25,920(n_{tp}) = 5,443,200$ deadmaps) are discussed. While the deadmap processing time ($\gtrsim$ 500 deadmaps/second on a single CPU and 16 GB of memory) is fast enough to evaluate even larger numbers of deadmaps (such as $\binom{20}{6} \cdot 25,920(n_{tp}) \simeq 1$B deadmaps), storing and loading more than approximately 5 million deadmaps requires some code adaptation, compressing the data, and writing, reading, and evaluating in a streamed fashion rather than loading all configurations into memory at once. Further speed-up is possible when running the model evaluation on a GPU and/or parallelising the evaluation.

6 wafers – constant per-wafer dead fraction

In this scenario, six wafers are simulated, each with a random but constant per-wafer dead fraction of $df_W = 5\%$. The total number of dead tiles is therefore the same across all wafers. All $n_{tp} = 25,920$ permutations are calculated and evaluated by the trained DNN predicting the DCAz loss in the range of $0.3 < p_T < 0.5$ GeV/c. Deadmap configurations are then ranked from best (least DCAz degradation) to worst (largest DCAz degradation). The per-layer dead fraction and outer region dead fraction are then evaluated. Layer L2 requires all segments of the wafer, therefore its dead fraction is a constant 5% (for the investigated case here). Layers L0 and L1, however, are cut out from wafers, and therefore the per-layer dead fraction varies.

The dead fraction per layer is shown in Figure 8.18a: The best-ranked deadmap configurations clearly favour an inner layer L0 with the lowest per-layer dead fraction. A similar, though less pronounced, trend is visible for L1, favouring lower dead fractions as well. As discussed above, the L2 dead fraction remains constant, since all available segments are used and the per-wafer dead fraction is fixed at $df_W = 5\%$. The corresponding dead tile fraction within the outer region (as defined in Figure 8.8) is shown in Figure 8.18b. The mean trend highlighted in red agrees with observations from deadmap rankings in Section 8.3.3, showing that better performance is achieved when a higher fraction of dead tiles is located in the outer regions. It should be noted, however, that we have only one fixed set of 6 wafers with defined dead tiles from which all deadmaps are calculated, rather than independently generated deadmaps for a full ITS3. This leads to a prior on the amount of dead tiles in the outer region (see also below), and constraints on the layer allocation. The 10 best and 10 worst deadmap configurations are shown in Figure 8.20a, illustrating the per-layer dead fractions df_{L0}, df_{L1}, df_{L2} and final DCAz loss. The predicted DCAz loss (for this set of 6 wafers) ranges from 2.7 % to 3.8 % for the best and worst configurations, respectively. This is both in agreement with the loss of individually simulated deadmaps (always for $0.3 < p_T < 0.5$ GeV/c, Appendix A.8c), and illustrates the range of DCAz loss depending on the chosen wafer-to-layer allocation and cutting selection for a constant wafer dead fraction of $df_W = 5\%$.



(a) Per-layer dead fraction.

(b) Outer region dead fraction.

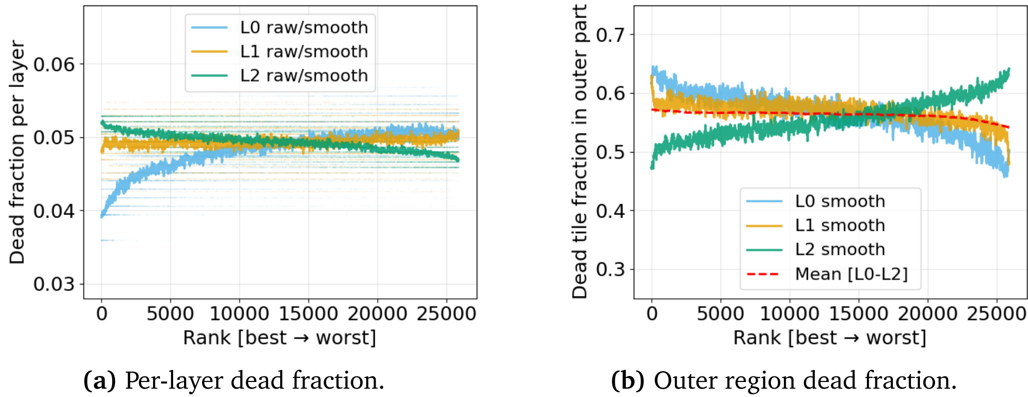
Figure 8.18: Deadmap configurations ranked by predicted DCAz loss for $0.3 < p_T < 0.5$ GeV/c and a constant per-wafer dead fraction of $df_W = 5\%$ across six wafers. (a) The per-layer dead fraction shows a clear favouring of a low L0 dead fraction for the lowest DCAz loss (best ranking). Raw data scatter markers show the discrete layer dead fractions depending on wafer-layer allocation and cutting pattern. The L2 dead fraction is constant, given the constant per-wafer dead fraction and no variability in cutting patterns for L2. (b) The dead fraction within the outer region shows the expected trend, with the best configurations showing a higher percentage of dead tiles in the outer part.

6 wafers – variable per-wafer dead fraction

A small variation in dead fraction between wafers is now introduced, and six wafers with wafer dead fractions of $df_W\,[\%] \in \{4.50, 4.75, 4.90, 5.10, 5.25, 5.50\}$ are simulated (mean $\langle df_W \rangle = 5\%$). Again, these dead tiles are fixed per-wafer (i.e. one draw of wafers), and all permutations are calculated, generating $n_{tp} = 25,920$ deadmaps. Deadmap configurations are evaluated with the DNN and ranked according to the predicted DCAz loss from least (best) to largest (worst). The corresponding dead fraction per layer is shown in Figure 8.19a. As expected, the best configurations minimise the dead fraction in the innermost layer L0. Layer L1 shows a similar but less pronounced trend, remaining largely flat. For L2, the trend is inverted: since all wafers must be allocated, prioritising lower dead fractions in L0 and L1 necessarily results in higher dead fractions in L2 for the best configurations. Similarly, Figure 8.19b shows that for the best configurations, higher dead tile fractions in the outer regions of L0 and L1 (where they have less impact on DCAz performance) are preferred. It should be noted that the mean outer-region dead tile fraction across the six wafers does not necessarily equal 50%; it depends on the specific random draw of

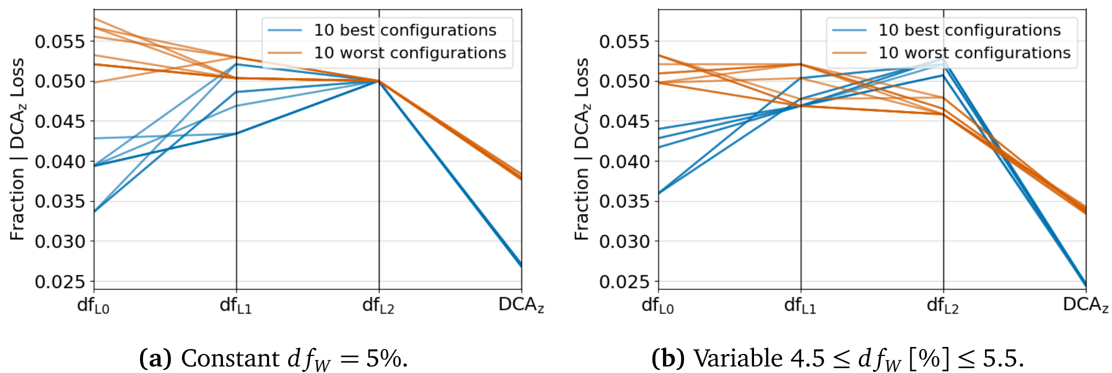
wafers. Across many such draws, however, a mean of approximately 50% is expected, consistent with the simulation results discussed in Section 8.3.3.

The 10 best and worst configurations are shown in Figure 8.20b, illustrating the per-layer dead fractions and the corresponding range of predicted best and worst DCAz losses.



(a) Per-layer dead fraction. **(b)** Outer region dead fraction.

Figure 8.19: Deadmap configurations ranked by predicted DCAz loss for $0.3 < p_T < 0.5$ GeV/c and a per-wafer dead fraction of $4.5 \leq df_W \leq 5.5\%$ for 6 wafers total. (a) The per-layer dead fraction clearly shows that configurations with the least DCAz loss (best ranking) favour a low dead fraction in L0. Given the 6-wafer constraint, the highest per-layer dead fraction is allocated to L2 as expected. (b) A higher L0 and L1 dead fraction within the outer detector part is favoured for best performance. Note that, e.g. compared to Figure 8.18b, the mean dead tile fraction in the outer part across all layers is > 50% – driven purely by the specific random draw of the six wafers used as the basis for all deadmap configurations in this scenario.



(a) Constant $df_W = 5\%$. **(b)** Variable $4.5 \leq df_W\,[\%] \leq 5.5$.

Figure 8.20: Parallel coordinates for the 10 best and 10 worst deadmap configurations for a per-wafer dead fraction of (a) $df_W = 5\%$ and (b) $4.5 \leq df_W \leq 5.5\%$ for 6 wafers and 25,920 configurations each. The per-layer dead fractions and final DCAz loss are shown. Although the mean per-wafer dead fraction is 5% in both cases, each scenario is based on a unique random assignment of dead tiles to wafers.

8 wafers – variable per-wafer dead fraction

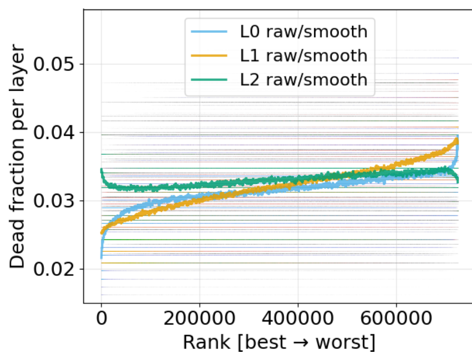
The ranking study is now extended to 8 wafers with per-wafer dead fractions of $df_W\,[\%] \in \{2.00, 2.25, 2.50, 3.00, 3.50, 4.00, 4.50, 5.00\}$, covering a larger range of wafer dead fractions. A single random draw of dead tile distributions is generated for each wafer, consistent with its assigned df_W. All $\binom{8}{6} \cdot 25,920(n_{tp}) = 725,760$ permutations and deadmaps are then calculated and ranked with the DNN according to predicted DCAz loss. The per-layer dead fraction, outer region dead fraction, and 10 best/worst configurations are shown in Figure 8.21a, Figure 8.21c, and Figure 8.21e, respectively. The combinatorial space is now increased, and not all wafers need to be allocated to an ITS3 layer (6 out of 8 wafers required). As in previous studies, the best-ranked configurations favour low per-layer dead fractions in L0 and L1, along with higher outer-region dead fractions in L0 and L1. The 10 best and 10 worst configurations further illustrate this trend, with the predicted DCAz loss spanning from approximately 1.1% (best) to 3.2% (worst). The two wafers with the highest per-wafer dead fractions are not allocated to any layer for the 10 best configurations. Interestingly, the allocation of the four wafers with the lowest df_W to L0 and L1 varies between configurations, as different cutting patterns can yield favourable ITS3 geometries despite similar wafer-level quality.

10 wafers – variable per-wafer dead fraction

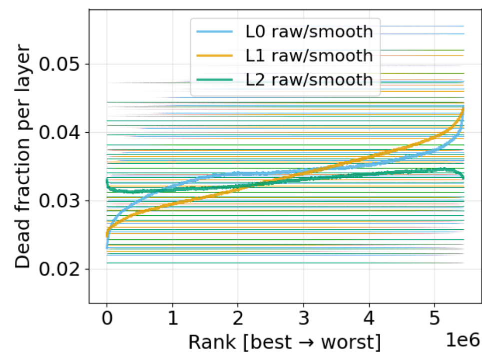
Finally, a set of 10 wafers with per-wafer dead fractions $df_W\,[\%] \in \{2.00, 2.25, 2.50, 2.75, 3.00, 3.50, 3.75, 4.00, 4.50, 5.00\}$ is simulated. All 5,443,200 corresponding deadmaps are computed and ranked based on the predicted DCAz loss. This demonstrates the feasibility of ranking configurations within a very large combinatorial space, with results consistent with previous observations.

The per-layer dead fraction, outer region dead fraction, and 10 best/worst configurations are shown in Figure 8.21b, Figure 8.21d, and Figure 8.21f, respectively. Trends again match previously observed characteristics for 6 and 8 wafers. Given the larger wafer-to-layer combinatorial space (6 out of 10 wafers), the L1 difference in per-layer dead fraction for the 10 best and worst configurations is now larger compared to

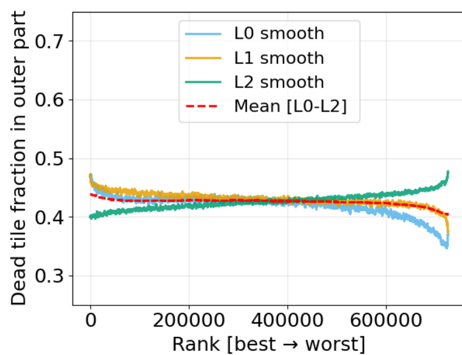
the 8-wafer case. Overall, for this specific draw of 10 wafers, the predicted DCAz loss spans from approximately 1.0% (best) to 3.4% (worst).



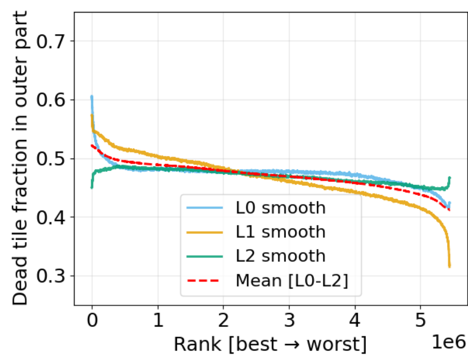
(a) Per-layer dead fraction. 8 wafers.



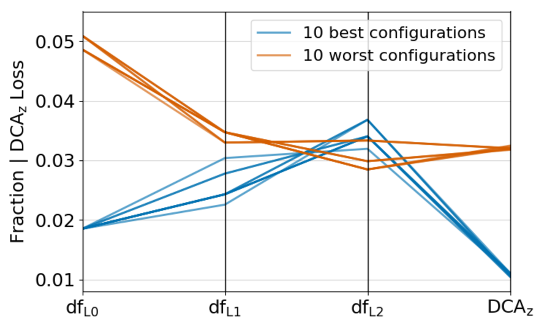
(b) Per-layer dead fraction. 10 wafers.



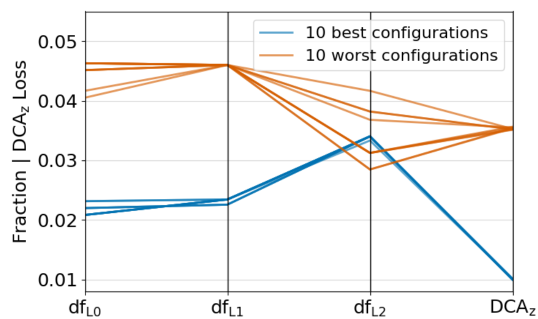
(c) Outer region dead fraction. 8 wafers.



(d) Outer region dead fraction. 10 wafers.



(e) 10 best/worst configurations. 8 wafers.



(f) 10 best/worst configurations. 10 wafers.

Figure 8.21: Deadmap configuration rankings for 8 wafers (a, c, e), and 10 wafers (b, d, f). The per-layer dead fraction, outer region dead fraction, and 10 best/worst configurations are shown. A total of 725,760 and 5,443,200 configurations are ranked for the set of 8 and 10 wafers, respectively. Discussion in text. Visible as horizontal lines in (a) and (b) are the raw scatter data of each configuration before smoothing.

Practical implications and model extension

Using an appropriately trained DNN to predict performance parameters, such as DCAz, from deadmaps proves to be a valuable tool for selecting and allocating wafers for the construction of the ITS3 barrel. Ultimately, extending or training an additional model for DCAxy performance, or adjusting the p_T range, should be considered. For example, a combined parameter $DCA_{combinedLoss} = \sqrt{DCA^2_{xyLoss} + DCA^2_{zLoss}}$ could be used as a ranking figure. Further model optimisation (see also Section 8.4.2), and additional training data are helpful. This includes considerations such as inefficient tiles (i.e. tiles that show reduced efficiency but are not switched off) or potential non-uniform distribution of failures at the wafer level. Overall, the qualitative two-condition approach (Section 8.3.4) aligns well with the findings of this study: the innermost layer L0 should have the fewest dead tiles, concentrated in the outer z-regions. Geometry optimisation is key to maximising detector performance, as demonstrated by the range of outcomes between the best and worst configurations illustrated in Figures 8.20, 8.21e, and 8.21f. A DNN-accelerated pre-selection of a set of best configurations, followed by full simulation runs for the final geometry choice, is one way to approach the task.

8.5 Effect of dead tiles on the Λ_c^+ reconstruction efficiency

To illustrate the effect of tracking efficiency loss on a physical observable, the Λ_c^+ baryon reconstruction is investigated, focusing on the impact of the ITS3 dead tiles on its reconstruction efficiency. With a mean lifetime of $\tau = (200 \pm 6)\ 10^{-15}$ s corresponding to $c\tau \simeq 60$ μm, the Λ_c^+ decays close to the primary vertex and well before the first tracking layer – relying on the ITS3's improved impact parameter resolution, as discussed in Section 2.4.2.

The three-prong decay $\Lambda_c^+ \rightarrow pK^-\pi^+$, with an (inclusive) branching ratio of $\Gamma_i/\Gamma = (6.23 \pm 0.33)\%$ [33], is studied, with all daughter particles tracked with the ITS. With the impact parameter resolution degradation approximately proportional

to the ITS3 dead fraction, this can be assumed a small effect on the measurement for the targeted $df \leq 2\%$[3]. Here, the effect of reduced reconstruction efficiency of the individual daughter tracks on the ability to reconstruct the Λ_c^+ is investigated. As a first-order approximation, the individual detection efficiency losses of the three daughter particles are combined to estimate the overall loss in Λ_c^+ reconstruction efficiency. This approach provides a fast estimate, re-using the data samples generated for the deadmap studies in Sections 8.2 and 8.3. It does, however, not consider an exact decay vertex or Λ_c^+ event reconstruction – which is needed for Λ_c^+ impact parameter resolution studies.

8.5.1 Daughter particle tracking loss

The loss in reconstruction efficiency is now calculated not only for pions as discussed in Section 8.3.2, but also for protons and kaons. The existing simulated data sets for both the reconstruction efficiency without dead tiles, and data sets for deadmaps with ITS3 dead fractions $df\,[\%] \in \{1, 2, 5, 10\}$ are re-evaluated in this study. The efficiency loss

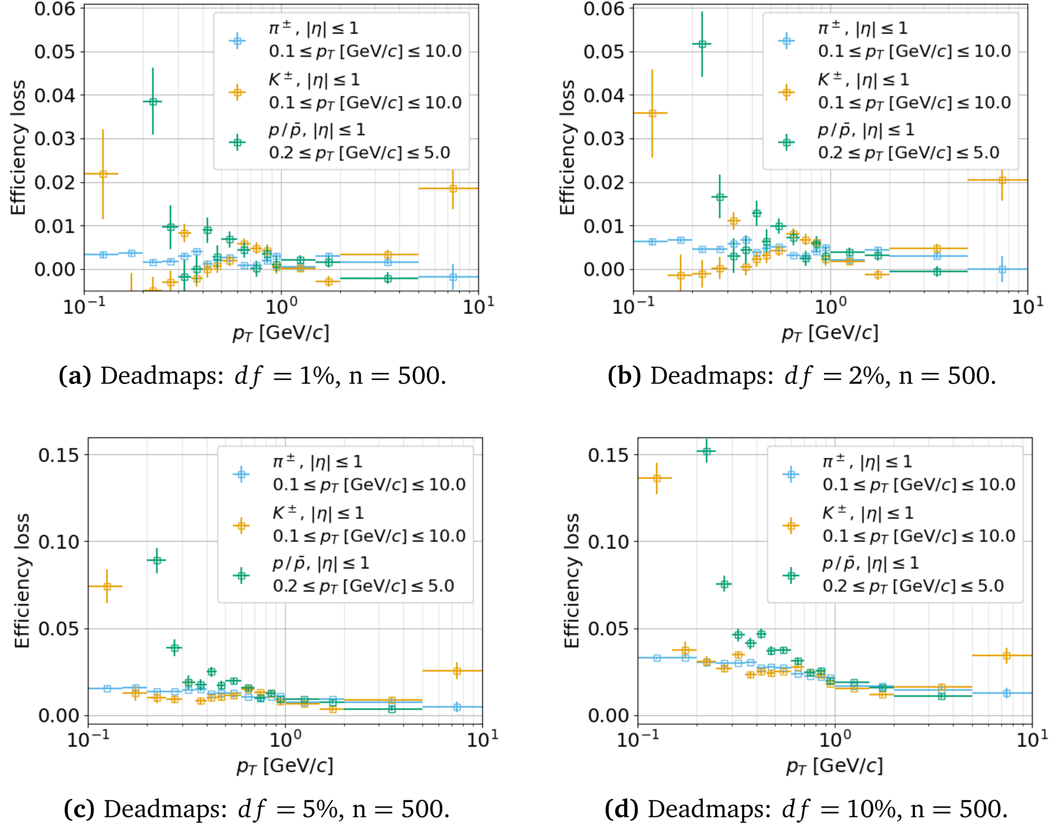
$$\epsilon_{loss}(p_T, \mathrm{X}) = \frac{eff_{full}(p_T, \mathrm{X}) - eff_{deadmap}(p_T, \mathrm{X})}{eff_{full}(p_T, \mathrm{X})}, \ \mathrm{X} \in \{p/\bar{p}, K^{\pm}, \pi^{\pm}\} \tag{8.12}$$

where $eff_{full}(p_T, \mathrm{X})$ denotes the particle-specific reconstruction efficiency with a fully operational ITS3 while $eff_{deadmap}(p_T, \mathrm{X})$ is the one with dead tiles. The resulting efficiency losses for all three particle species are shown in Figure 8.22. The proton p_T range is reduced due to the limited number of events. This, however, does not affect the study, since a lower cut at $p_T > 0.3$ GeV/c for all daughter particles is performed, as discussed below.

8.5.2 Event generation and kinematics

All signal events are generated using a standalone MC using PYTHIA 8.311 [74] for the decay chain simulation. The parent Λ_c^+ p_T is sampled from a central FONLL [180]

[3]Considering the factor two improvement in impact parameter resolution with the inclusion of the ITS3 over the current ITS2 configuration, yielding a factor four and ten improvement in significance and S/B, respectively (cf. Figure 2.12).



(a) Deadmaps: $df = 1\%$, n = 500.

(b) Deadmaps: $df = 2\%$, n = 500.

(c) Deadmaps: $df = 5\%$, n = 500.

(d) Deadmaps: $df = 10\%$, n = 500.

Figure 8.22: Efficieny loss for $p/\bar{p}, K^\pm, \pi^\pm$ at simulated ITS3 dead fractions of (a) 1%, (b) 2%, (c) 5%, (d) 10%.

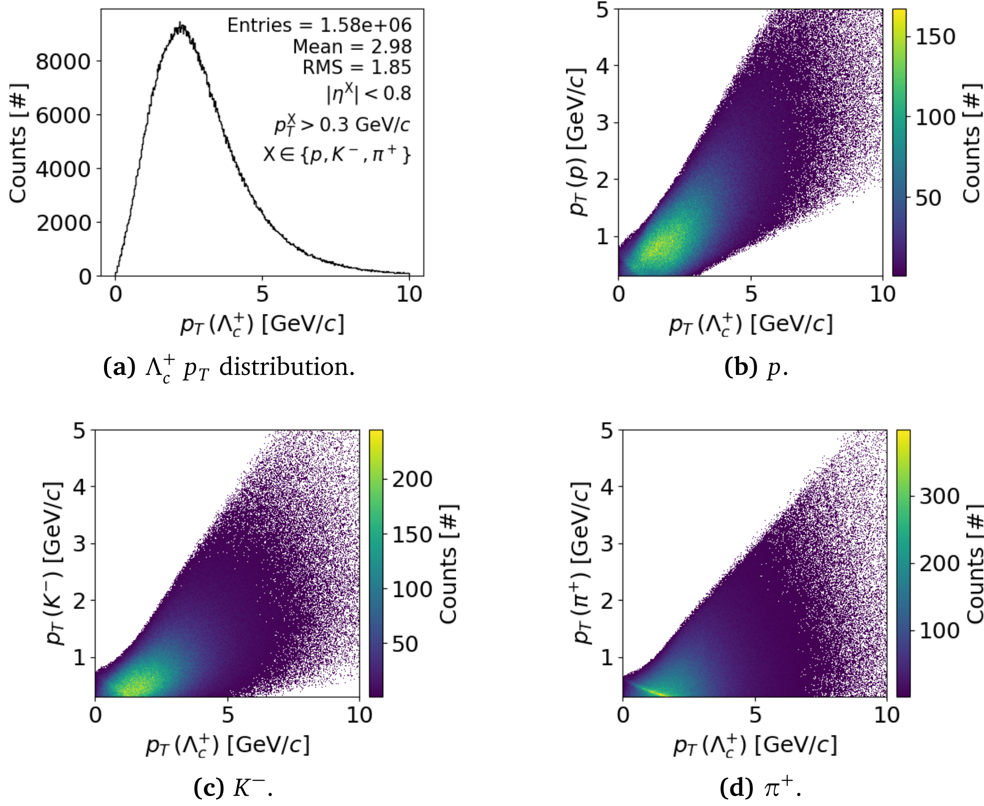
cross section prediction for prompt production at $\sqrt{s_{pp}} = 5.02$ TeV [4]. The full four-vector of the Λ_c^+ is calculated and inserted into PYTHIA at the nominal interaction point. Only Λ_c^+ decay channels with the $pK^-\pi^+$ final state are enabled (see also Section 2.4.2). An analysis-motivated kinematic filter is applied to all three daughter tracks simultaneously, requiring[5]

$$p_T^X > 0.3\ \text{GeV}/c, \ \ |\eta^X| < 0.8, \ \ X \in \{p, K^-, \pi^+\}. \tag{8.13}$$

The distribution of selected Λ_c^+ events is shown in Figure 8.23a. The corresponding daughter particle momenta distributions are shown in Figure 8.23b, Figure 8.23c, and Figure 8.23d, for protons, kaons, and pions, respectively. The Dalitz plot in Figure 2.13 illustrates the kinematic phase space.

[4]The p_T distribution itself is the FONLL D^0 prediction multiplied by the measured Λ_c^+/D^0 p_T differential ratio [82], given that FONLL does not provide baryon predictions.

[5]Limiting – for real measurements – the large number of 3-particle combinations below $p_T <$ 0.3 GeV/c, and ensuring reliable PID performance of the TPC at $|\eta| < 0.8$.



(a) Λ_c^+ p_T distribution. **(b)** p. **(c)** K^-. **(d)** π^+.

Figure 8.23: Transverse momenta distributions of selected Λ_c^+ candidates (a), and corresponding daughter particles (b–d).

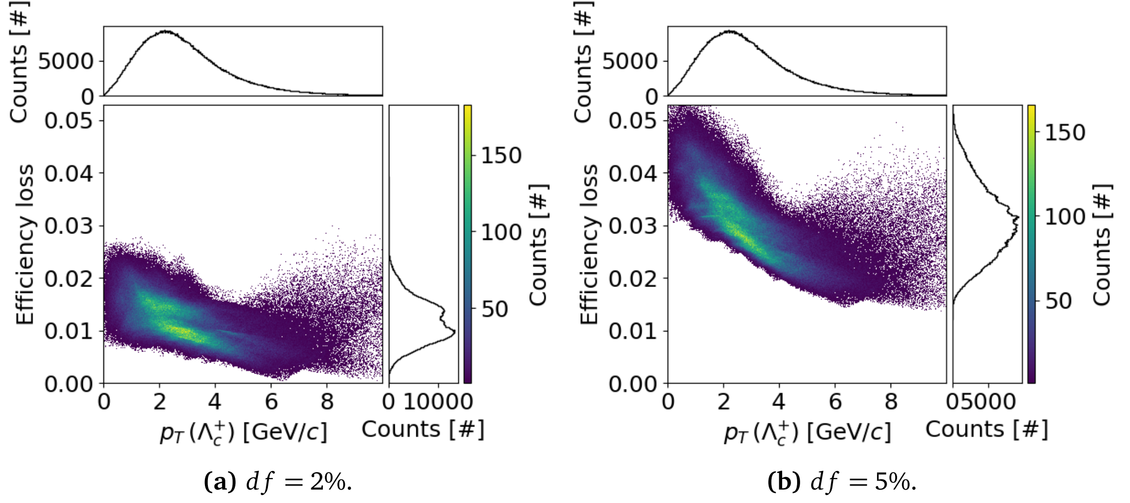
A small fraction of $\simeq 0.5\%$ events with a proton daughter $p_T(p) > 5\text{GeV}/c$ are discarded as the proton daughter loss simulation suffers from an insufficient number of events in this high-p_T region.

8.5.3 Λ_c^+ efficiency loss

The Λ_c^+ tracking efficiency loss is estimated for simulated ITS3 dead fractions $df\,[\%] \in \{1, 2, 5, 10\}$. In a first-order approximation, the reconstruction efficiency losses of the three daughter particles are assumed to be independent. The resulting Λ_c^+ reconstruction efficiency loss is then calculated per event as

$$\epsilon_{loss}^{\Lambda_c^+} = 1 - \left[(1 - \epsilon_{loss}^{p}) \cdot (1 - \epsilon_{loss}^{K^-}) \cdot (1 - \epsilon_{loss}^{\pi^+})\right]. \tag{8.14}$$

For each event and each Λ_c^+ daughter particle, the reconstruction efficiency loss for each daughter as a function of the daughter's transverse momentum and a given ITS3



(a) $df = 2\%$. **(b)** $df = 5\%$.

Figure 8.24: 2D distribution of tracking efficiency loss versus Λ_c^+ transverse momentum for dead fractions of (a) 2% and (b) 5%.

dead fraction is extracted from the corresponding simulation (see Figure 8.22). The efficiency loss and uncertainty for momenta in between bins are linearly interpolated.

The resulting Λ_c^+ efficiency loss for simulated ITS3 dead fractions of 2% and 5% are shown as 2D histograms in Figure 8.24a and Figure 8.24b, respectively. The p_T-binned mean efficiency loss profile is given in Figure 8.25 for ITS3 $df\ [\%] \in \{1, 2, 5, 10\}$. The daughter efficiency loss is propagated assuming the three daughter particles are uncorrelated:

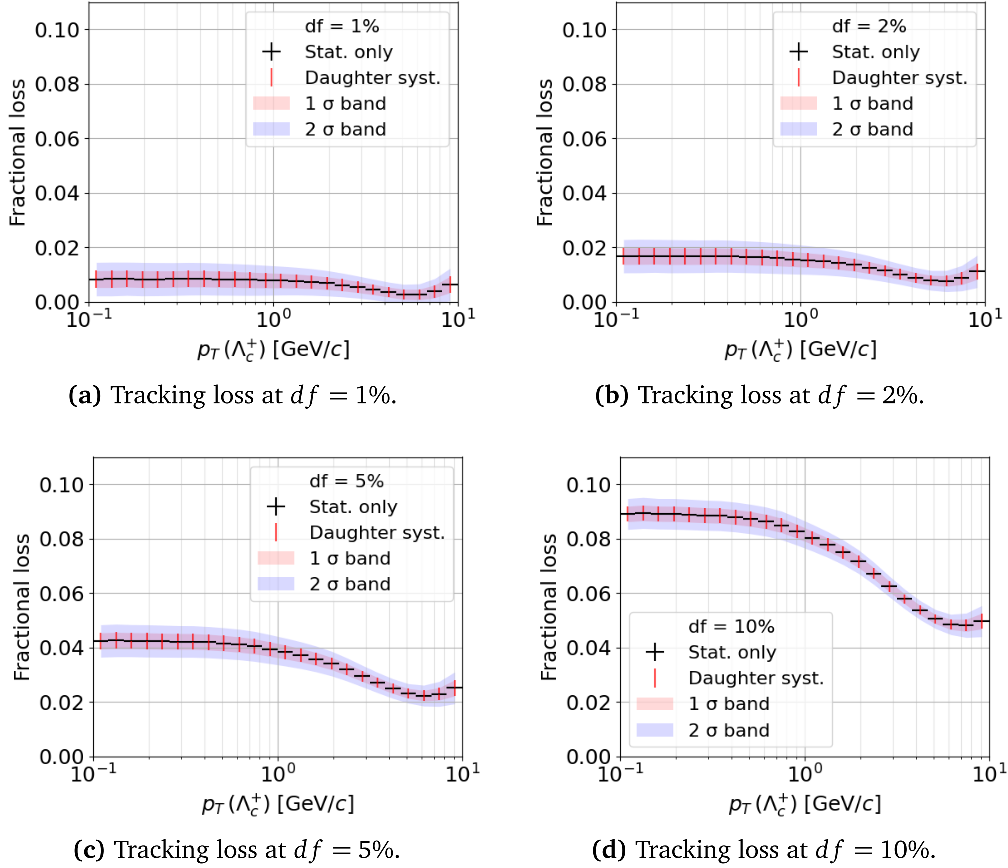
$$\sigma^2_{\epsilon_{loss}^{\Lambda_c^+}} = \sum_{\mathrm{X}} \left(\frac{\partial \epsilon_{loss}^{\Lambda_c^+}}{\partial \epsilon_{loss}^{\mathrm{X}}} \right)^2 \sigma^2_{\mathrm{X}}\,, \quad \mathrm{X} \in \{p, K^-, \pi^+\}\,, \tag{8.15}$$

where σ_{X} is the efficiency loss uncertainty of particle X. The statistical uncertainty per p_T-bin is computed as the standard error of the mean $\sigma_{\mathrm{stat}}(p_T) = S(p_T)/\sqrt{N}$. The total uncertainty per p_T-bin is then the quadrature sum

$$\sigma_{tot}(p_T) = \sqrt{\sigma^2_{stat}(p_T) + \sigma^2_{sys}(p_T)} \tag{8.16}$$

where the systematic component is given by:

$$\sigma^2_{sys}(p_T) = \frac{1}{N} \sum_{k=1}^{N} \sigma^{(k)}_{\epsilon_{loss}^{\Lambda_c^+}}(p_T) \tag{8.17}$$



(a) Tracking loss at $df = 1\%$.

(b) Tracking loss at $df = 2\%$.

(c) Tracking loss at $df = 5\%$.

(d) Tracking loss at $df = 10\%$.

Figure 8.25: Λ_c^+ tracking loss at simulated ITS3 dead fractions of (a) 1%, (b) 2%, (c) 5%, (d) 10%.

The 1 σ and 2 σ bands represent $\langle\epsilon_{loss}^{\Lambda_c^+}(p_T)\rangle \pm \sigma_{tot}(p_T)$ and $\langle\epsilon_{loss}^{\Lambda_c^+}(p_T)\rangle \pm 2\sigma_{tot}(p_T)$, respectively.

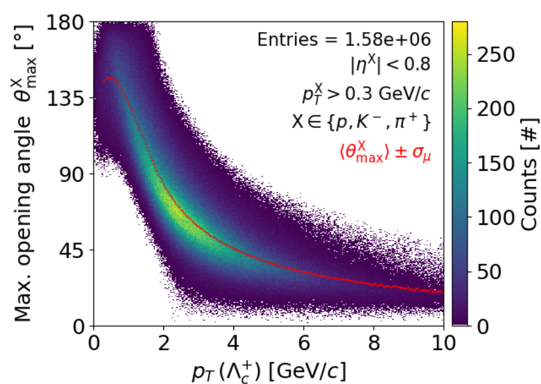
For momenta $p_T(\Lambda_c^+) \lesssim 1$ GeV/c, the loss in Λ_c^+ tracking efficiency is approximately proportional to the ITS3 dead fraction. At higher transverse momenta (e.g. mean $\langle\Lambda_c^+(p_T)\rangle = 2.98$ GeV/c), the efficiency loss drops to about 60% of the dead fraction percentage. ITS3 dead fractions of below 5% can therefore be considered acceptable from an efficiency-loss viewpoint for the Λ_c^+ reconstruction. These estimates are conservative (see below). For a similarly sized effect for both signal (S) and background (B), a reduction of 5% efficiency leads to a small increase in statistical uncertainty of $1/\sqrt{1-0.05} - 1 \simeq 2.6\%$.

Boosted decay consideration

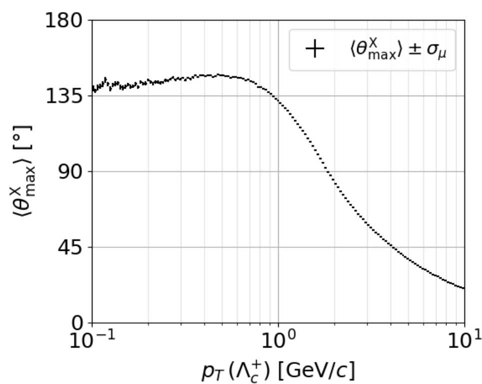
At increasing $p_T(\Lambda_c^+)$, the tracks of the three daughter particles $\mathrm{X} \in \{p, K^-, \pi^+\}$ are expected to become increasingly correlated. At high collimation, the daughters traverse similar regions of the detector, and consequently, the efficiency loss estimated by assuming independent daughter tracks tends to overestimate the true loss. For any two daughter-momentum 3-vectors $\mathbf{p}_i$, $\mathbf{p}_j$ the opening angle θ^{ij} is calculated as $\theta^{ij} = arccos\left(\frac{\mathbf{p}_i \cdot \mathbf{p}_j}{|\mathbf{p}_i||\mathbf{p}_j|}\right)$. All pair-wise angles $\theta^{pK^-}, \theta^{p\pi^+}, \theta^{K^-\pi^+}$ are calculated and the maximum opening angle of the decay extracted as $\theta^{\mathrm{X}}_{\mathrm{max}} = max\{\theta^{pK^-}, \theta^{p\pi^+}, \theta^{K^-\pi^+}\}$. The resulting 2D histogram illustrating the p_T dependent maximum opening angle is shown in Figure 8.26a with the mean maximum opening angle on a logarithmic p_T scale shown in Figure 8.26b. Above momenta of $O(1\ \mathrm{GeV}/c)$ all three decay products are increasingly collimated with the maximum opening angle dropping to below 25° at $p_T = 10\ \mathrm{GeV}/c$.

Similarly, $\theta^{\mathrm{X}}_{\mathrm{min}} = min\{\theta^{pK^-}, \theta^{p\pi^+}, \theta^{K^-\pi^+}\}$, the minimum opening angle between any two daughters, is shown in Figure 8.26c and Figure 8.26d as a 2D histogram and for the mean minimum opening angle $\langle\theta^{\mathrm{X}}_{\mathrm{min}}\rangle$, respectively. Any two daughters show a higher mean collimation, dropping to a mean minimum pairwise opening angle of $\lesssim 10°$ at $p_T = 10\ \mathrm{GeV}/c$.

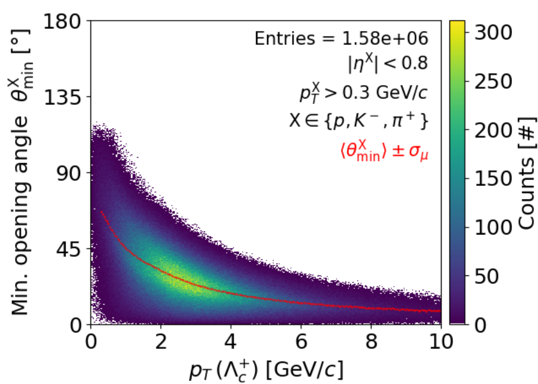
Thus, the Λ_c^+ three-prong decay is increasingly boosted and collimated for transverse momenta $\gtrsim 1\ \mathrm{GeV}/c$. Nevertheless, the independent daughter track approach for tracking loss estimation remains reasonable. For the ITS3 layer L0 and a tile arc length of approximately 8 mm, the r—φ opening angle is about 24° (worst case). While the decay is boosted and decay products traverse similar sections of the ITS, and are therefore not fully uncorrelated, the independent daughter reconstruction efficiency loss (Equation (8.14)) for dead tiles in the ITS3 still holds for $p_T(\Lambda_c^+) \lesssim 10\,\mathrm{GeV}/c$. The estimates for the efficiency loss from ITS3 dead fractions are therefore conservative.



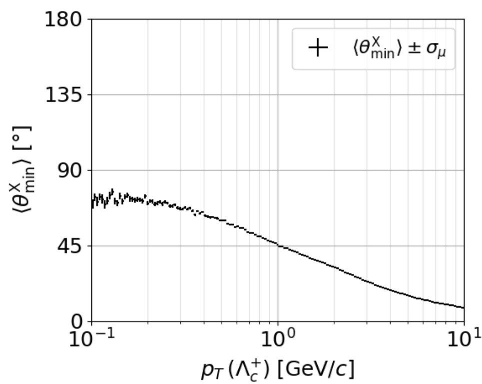
(a) Maximum opening angle distribution.



(b) Mean maximum opening angle.



(c) Minimum opening angle distribution.



(d) Mean minimum opening angle.

Figure 8.26: Maximum opening angle of all three daughter particles: (a) per event, (b) mean. Minimum opening angle of any two daughter pairs: (c) per event, (d) mean.

9 Conclusions and Outlook

This work advances a novel vertex-detector concept that will be realised for the first time in ITS3. The ultra-low mass, air-cooled, high spatial resolution device – bent into a cylindrical shape around the beam pipe – will deliver unprecedented impact parameter resolution and tracking efficiencies down to transverse momenta well below 1 GeV/c for ALICE. It will be a new tool for future particle and nuclear physics experiments and the first of its kind.

To validate the stitching technique required to manufacture the ITS3 wafer-scale sensors, far exceeding design reticle dimensions, the prototype Monolithic Stitched Sensor MOSS was studied in this work. Custom handling, mounting, and interconnection tools and procedures were developed. High reliability was essential, given the limited number of test objects. In total, 82 MOSS chips were successfully picked, mounted, and assembled, achieving a 98% success rate. The developed procedure was adapted for additional test structures not discussed here.

For an in-depth characterisation of the MOSS sensor, a careful approach was adopted, and dedicated impedance and power-ramp setups were built, including thermal camera imaging for localising heat signatures. A complex, FPGA-based measurement system, including a versatile software suite, was developed for functional chip characterisation.

Early measurements with both the impedance and power ramping setups revealed concerning on-chip short-circuit faults on all wafers. Understanding the root cause was critical. Although 89% of short structures could be opened by applying a sufficiently large but moderate current, allowing the sensors to function, this workaround is clearly not viable. A new analysis method was developed, correlating impedance measurements, thermal camera images, and power ramping data with the chip design files, thereby allowing the formation of a hypothesis on the short fault mechanism. Statistical tests showed excellent agreement with the hypothesis. For further confirmation, Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) was employed, clearly showing the short structure as hypothesised in two separate samples, and additionally validating the accuracy of the developed failure analysis method. In dialogue with the foundry, a processing issue was identified in the affected region of the on-chip metal stack, which led to the observed faults. Future devices will employ a new metal stack composition, and together with an adjusted design strategy, these failures are expected to be eliminated. Functional chip characterisation returned a yield compatible with ITS3 requirements, confirming that the planned wafer production is expected to meet sensor specifications.

The ITS3 R&D programme has already achieved significant milestones, including silicon bending, air-cooling integration, and the 65 nm CMOS process validation. With the MOSS sensor, the first step in confirming the feasibility of using stitching for high-energy physics MAPS has now been taken. At the time of writing, the next sensor is being submitted for fabrication. It will integrate the full functionality required for the final ITS3 sensors, building on the results of the 65 nm and MOSS characterisation campaigns. It represents the final submission prior to the production of sensors intended for installation in ALICE. Impedance and power ramp-up measurements, as developed in this thesis, will be integrated as an initial characterisation step.

A non-zero failure rate must be assumed on the final sensors, and design choices were made to account for this, allowing to switch off individual sub-structures (tiles) on-chip. The effect of dead tiles on the impact parameter resolution and track reconstruction efficiency was studied. For $\pi^{\pm}$ in the momentum range $0.1 <$

$p_T < 10$ GeV/c, the impact-parameter resolution degrades roughly in proportion to the ITS3 dead fraction, while tracking-efficiency losses are less than half as large. The functional wafer map will be determined with wafer probing, before cutting sensors and allocating them to the ITS3 layers. Ranking the simulated imperfect layer configurations according to impact parameter resolution validated the strategy of assigning the best-performing sensors to the innermost layers of ITS3. Given the number of wafers, the combinatorial problem of cutting and assigning sensor planes becomes computationally prohibitive for exhaustive simulation. The feasibility of employing a deep neural network to optimise the geometrical arrangement of ITS3 sensors was explored. The network yielded results consistent with full simulations and provides a promising foundation for further studies aimed at navigating the vast configuration space.

Large-area, wafer-scale, flexible MAPS are on track for use in cutting-edge vertex and tracking detectors for current and future experiments. Beyond their immediate application in ITS3, similar performance requirements are anticipated for the proposed ALICE 3 experiment, the ePIC detector at the Electron-Ion Collider (EIC), and potential future Higgs factories such as the FCC-ee – all of which are likely to build upon the technologies and concepts developed for ITS3. Further R&D and exploration of new ideas like the embedding of MAPS into flexible printed circuits [181], or integrating microchannel cooling on-chip (demonstrated for small-scale MAPS [182]) are exciting opportunities to push the boundaries of detector performance further, and *make measurable what is not so*.

Alles Seiende wollt ihr erst denkbar machen: denn ihr zweifelt mit gutem Misstrauen ob es schon denkbar ist. ... Und dies Geheimnis redete das Leben selber zu mir: »Siehe«, sprach es, »ich bin das, was sich immer selber überwinden muss.«

– Friedrich Nietzsche, Zarathustra

Appendices

A
Supplementary Figures

Supplementary figures to the main text are given in this Appendix in the order of occurrence.

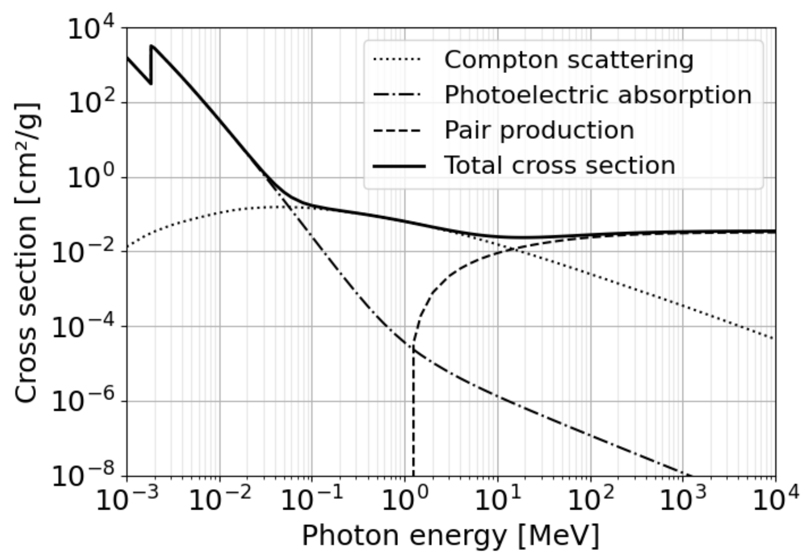


Figure A.1: Photon cross section in silicon. Data from [89].

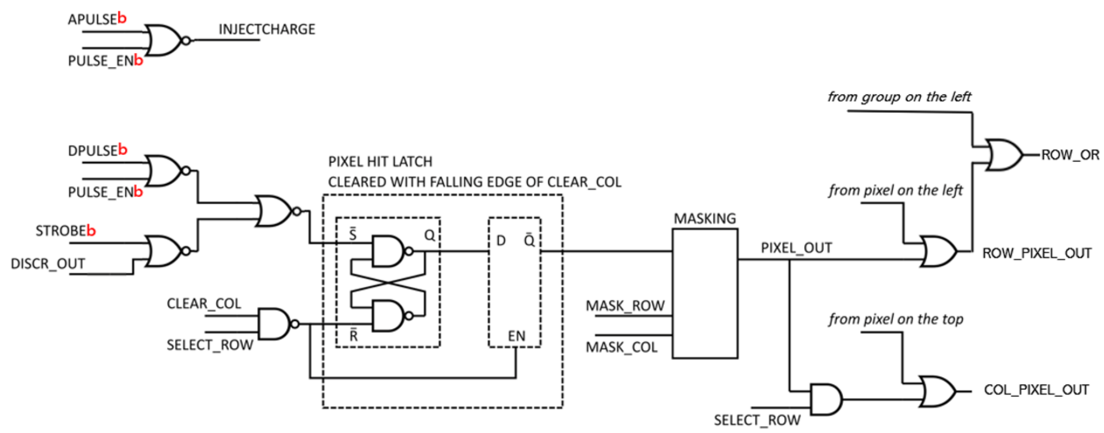


Figure A.2: MOSS chip digital pixel logic, including masking functionality.

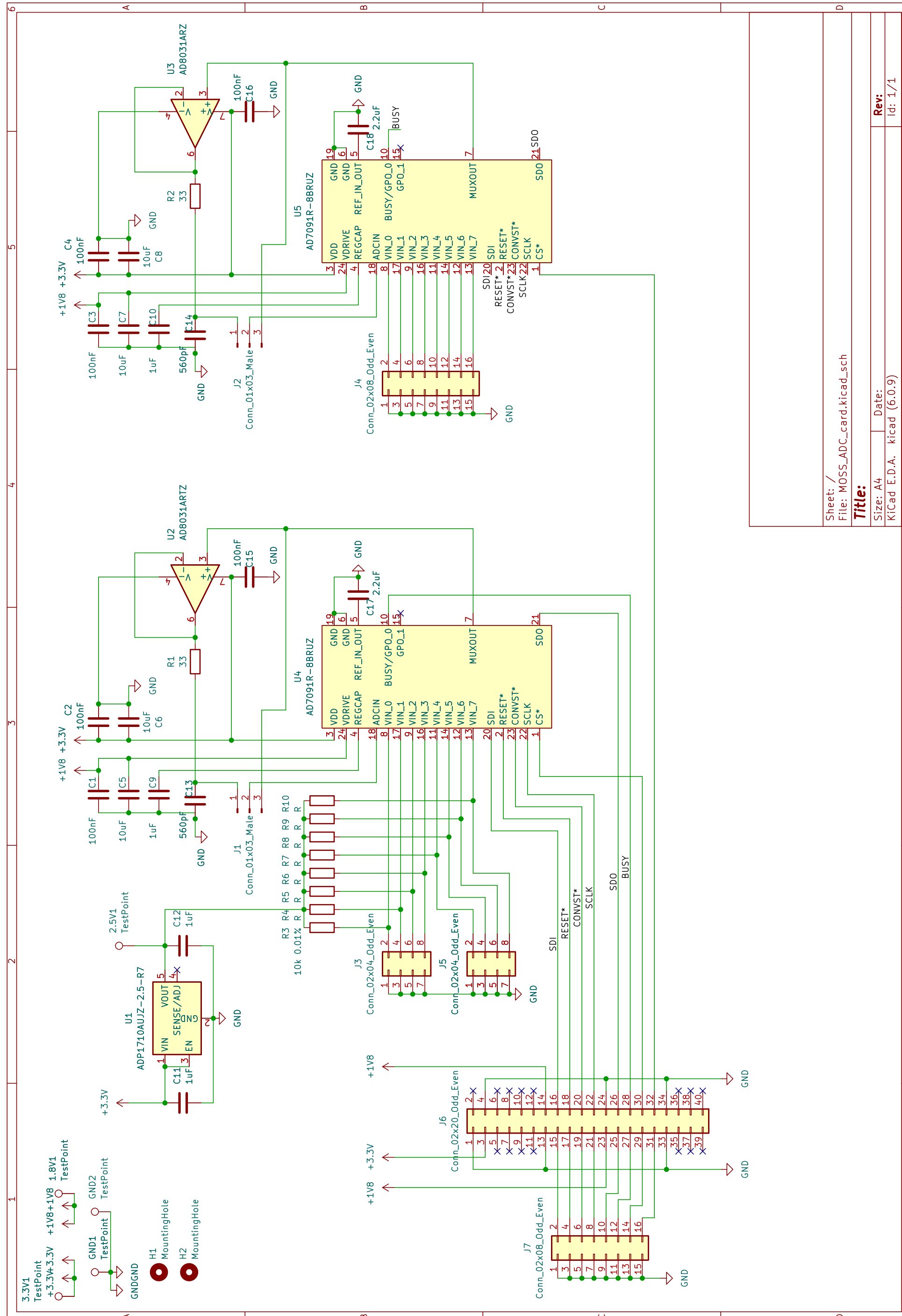


Figure A.3: Schematic of the NTC-ADC board.

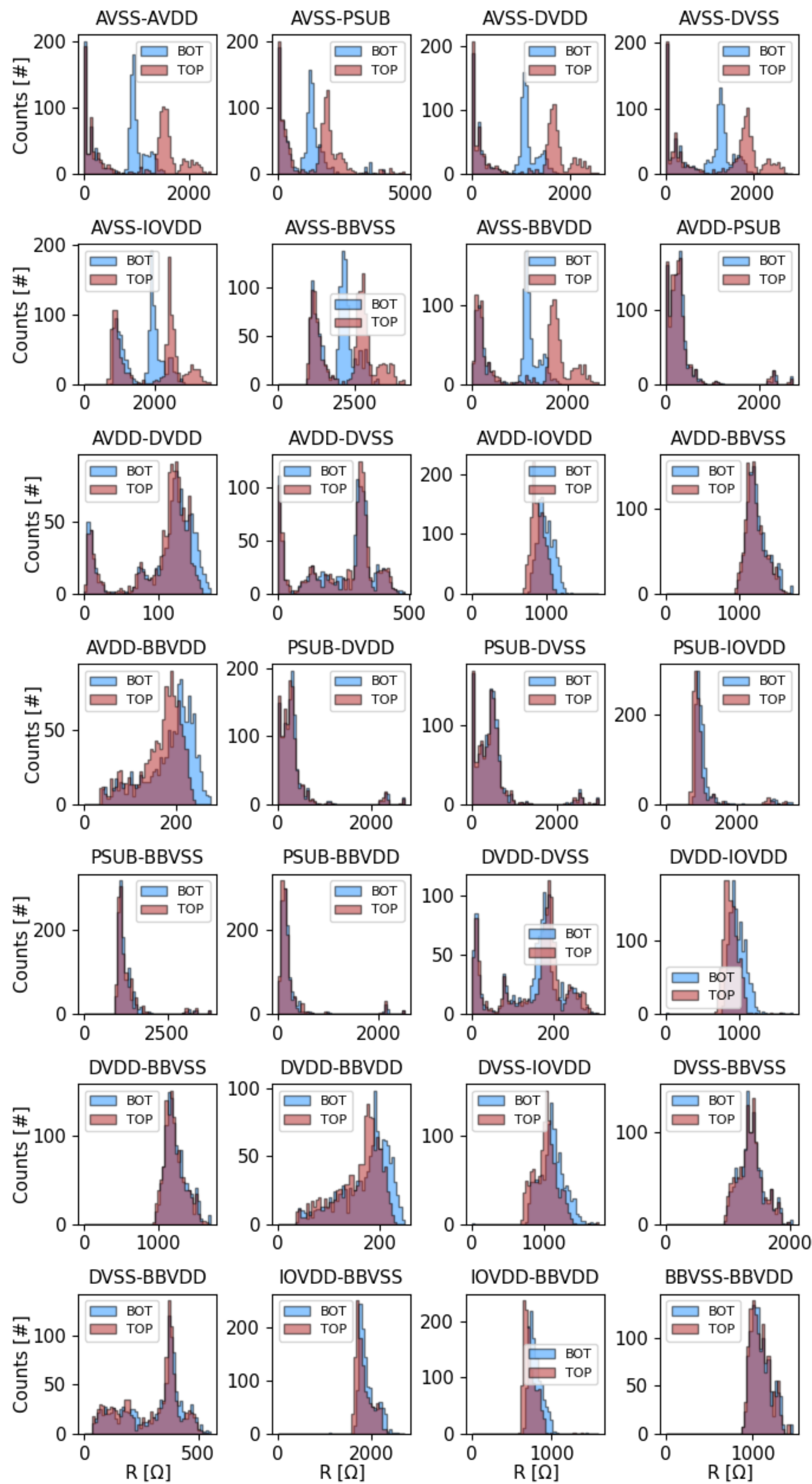
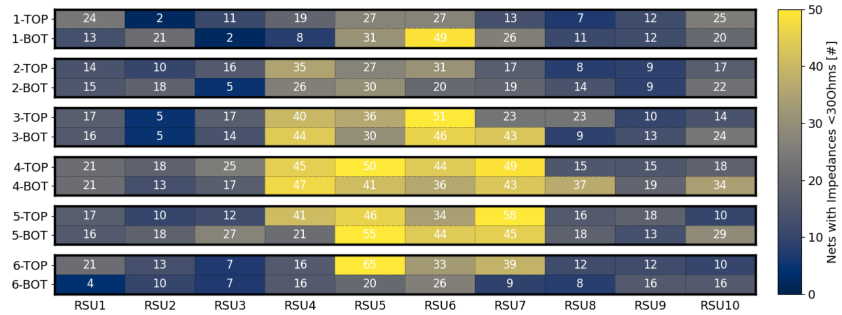
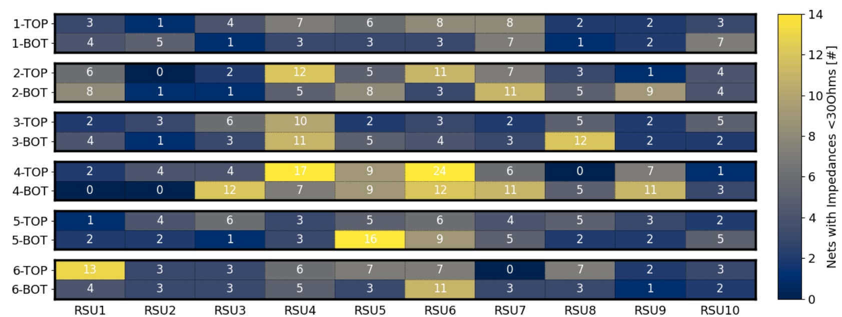


Figure A.4: Impedances for all 28 power net pair combinations split in TOP/BOT HUs before powering. The distinct distributions for top and bottom units when the analogue domains (AVDD, AVSS) are involved stem from the different pixel matrices in top and bottom, respectively.



(a) Odd-numbered wafers.



(b) Even-numbered wafers.

Figure A.5: Wafer maps of short distributions split into (a) odd and (b) even numbered wafers. Gaps between the MOSS sensor positions are to scale, as manufactured on-wafer.

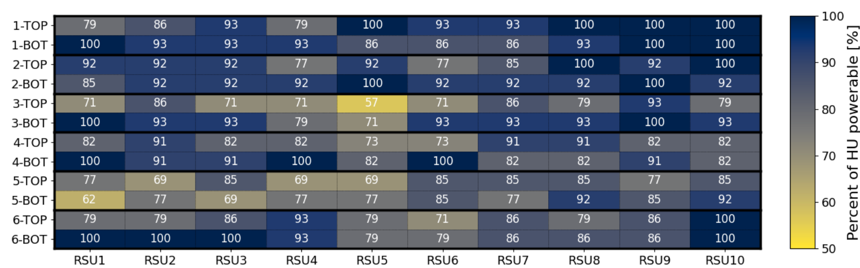
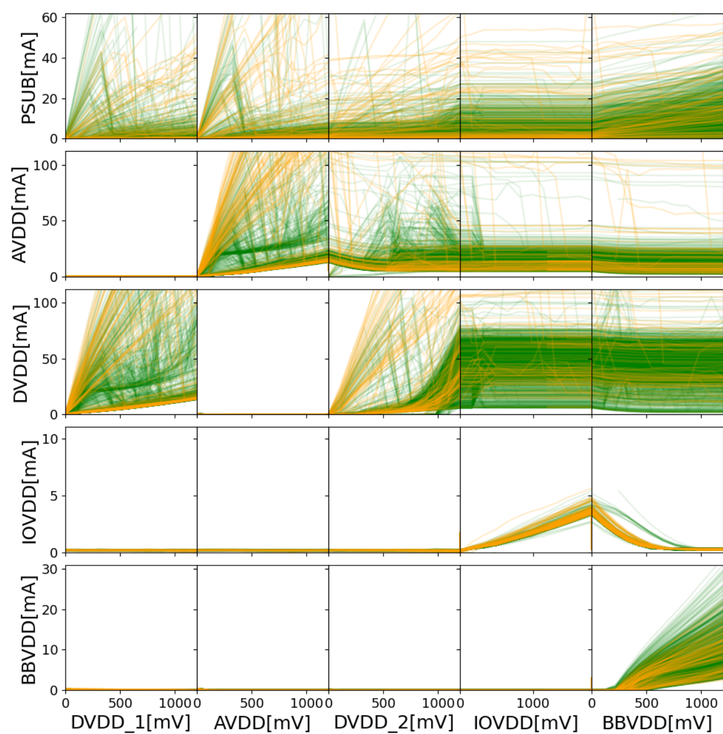
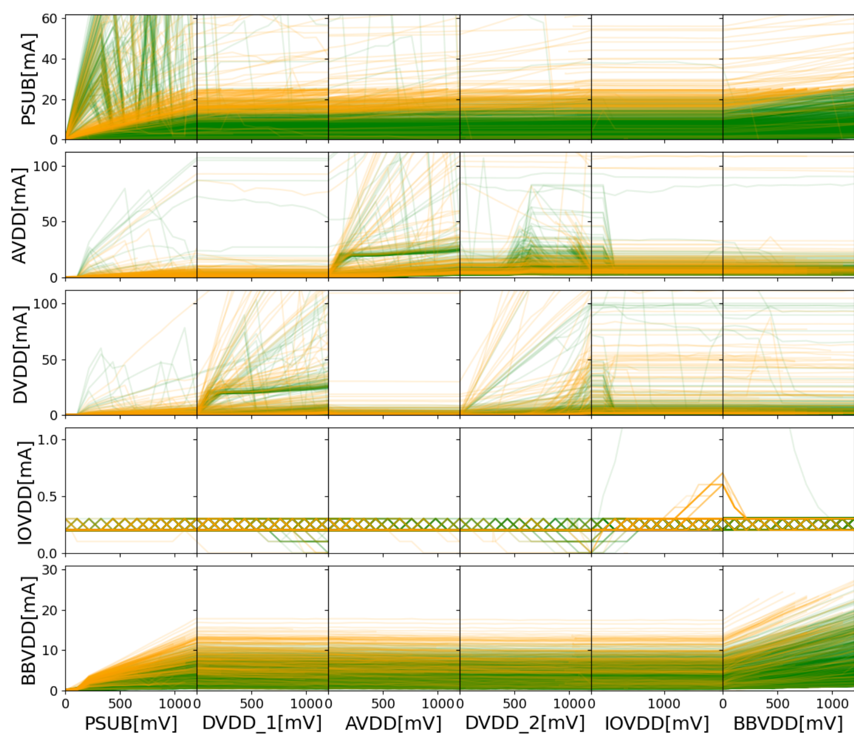


Figure A.6: Fraction of powerable HUs with extended PSUB limit as used in the test system.

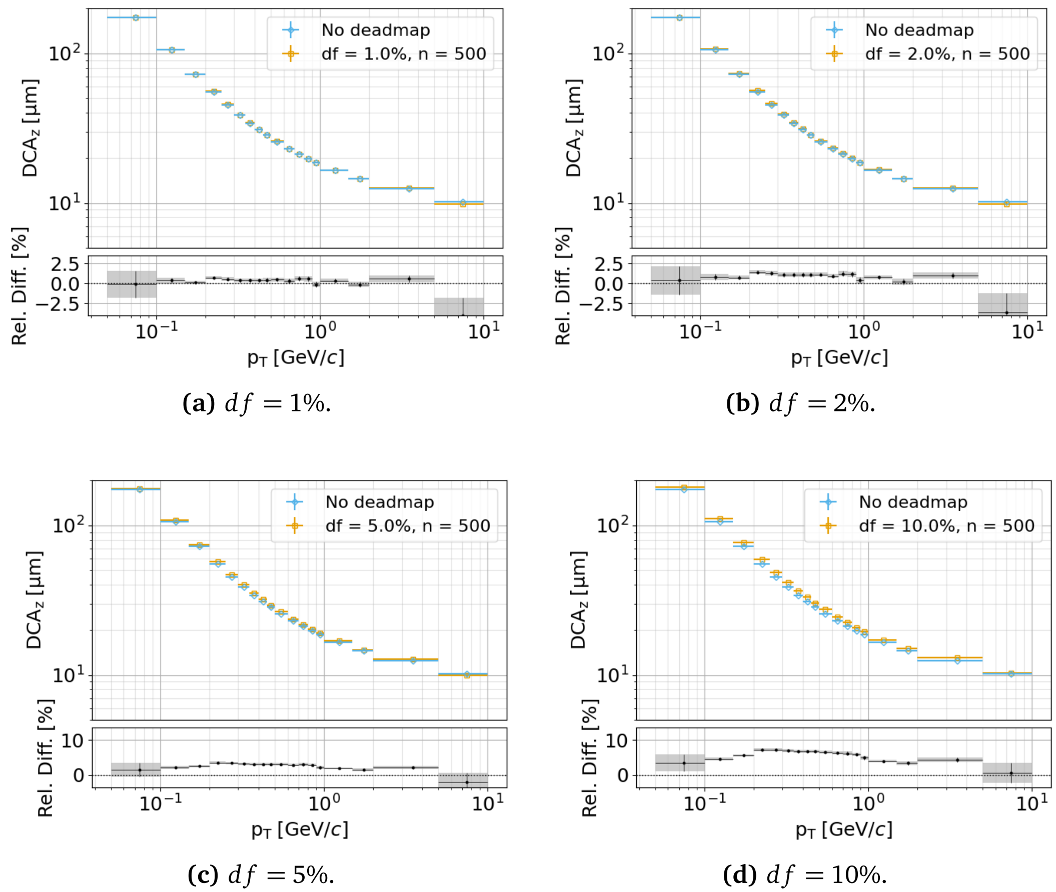


(a) PSUB = 0 V.



(b) PSUB = –1.2 V.

Figure A.7: Power ramp-up shapes for (a) PSUB = 0 V, and (b) PSUB = –1.2 V. Ramps failing the ramp-up at any stage are shown in orange.



(a) $df = 1\%$.

(b) $df = 2\%$.

(c) $df = 5\%$.

(d) $df = 10\%$.

Figure A.8: Pointing resolution loss in the longitudinal plane for simulated ITS3 dead fractions of 1% (a), 2% (b), 5% (c), 10% (d).

B
Extended Yield Model with Clustering Factor

The main text derives the wafer lot extrapolation from a Poisson assumption of region failures. The functionally failed regions per MOSS sensor (excluding limitations due to chip architecture constraints) are shown in Figure B.1.

The corresponding Poisson ($f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$) and Negative-Binomial (NB) distributions are overlaid. The NB probability density, as a Poisson-gamma mixture model, is defined as [183, 184]:

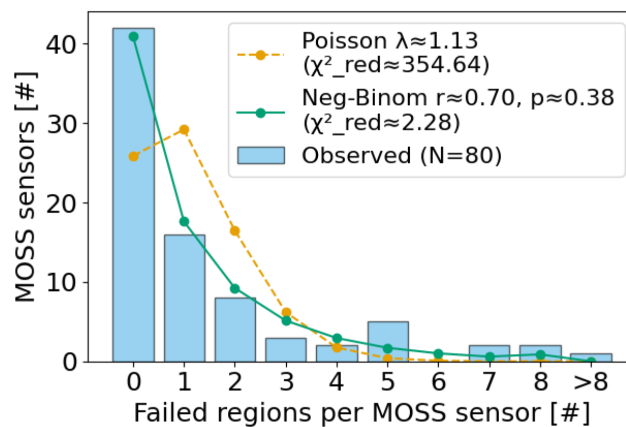


Figure B.1: Functionally failed number of regions per MOSS sensor (there are 80 regions per MOSS sensor) for 80 chips. As not all HUs were tested for every MOSS, the number is scaled to the nearest integer where needed.

$$f(k;r,p) = \frac{\Gamma(r+k)}{k!\,\Gamma(r)}(1-p)^k p^r\,, \tag{B.1}$$

with

- $r = C =$ dispersion, or 'clustering factor',
- $p =$ success-probability,
- $k =$ trials,
- $\mu = r\,(1-p)/p$, and
- $\sigma^2 = \mu + \mu^2/r$,

where μ and σ^2 are the sample mean and variance, respectively. Parameters $r = C$ and p are directly derived from the calculated sample mean and variance. For $r \to \infty$, the distribution converges to a Poisson distribution, whereas for small r the variance of the NB distribution exceeds the mean. It is an appropriate choice when a gamma shape can describe the overdispersion in an otherwise Poisson model.

The functional per-region yield Y_r for $N = 80$ regions per sensor and a mean μ of non-working regions per sensor is then $Y_r = 1 - \mu/N$ – equivalent to the directly observed $Y_r\langle\text{MOSS}\rangle = 98.3\%$, as per Section 6.6. While the Poisson model provides a proper estimate, the NB model better describes the observed distribution, allowing for 'overdispersion' of the Poisson model. For the NB model, the clustering factor $r = C$ is introduced as [110, 115]:

$$Y_r(A_r) = \left(1 + \frac{A_r D_M}{C}\right)^{-C} \to \lim_{C\to\infty}\left(1 + \frac{A_r D_M}{C}\right)^{-C} = \mathrm{e}^{-A_r D_M}. \tag{B.2}$$

For $C \to \infty$, the form in Section 6.6 is reproduced. With an extracted clustering factor $C = 0.7$ (see Figure B.1), the equivalent NB defect density $D_M \to D_M^{NB}$ becomes:

$$D_M^{NB} = \frac{C}{A_r}\left(Y_r^{-1/C} - 1\right). \tag{B.3}$$

The corresponding ITS3 sensor-scaled per-tile pass probability $p_{tile} \to p_{tile}^{NB}$ becomes:

$$p_{tile}^{NB} = \left(1 + \frac{A_t D_M^{NB}}{C}\right)^{-C} . \tag{B.4}$$

The wafer-pass probability (6.11) and lot size calculation (6.12) follow the procedure outlined in Section 6.6, and the resulting wafer-lot size diagram for ITS3 layer extrapolation is shown in Figure B.2. Ultimately, the difference between the Poisson and NB models is negligible for estimating wafer lot sizes in this instance (e.g. at a MOSS functional yield of 98.5%, the estimated required lot size for a $\leq 1.0\%$ ITS3 dead fraction increases from 25 to 26), where other effects, such as wafer-to-wafer variation, lot-to-lot variation (unknown), and new chip features with unknown yield, play a much larger role.

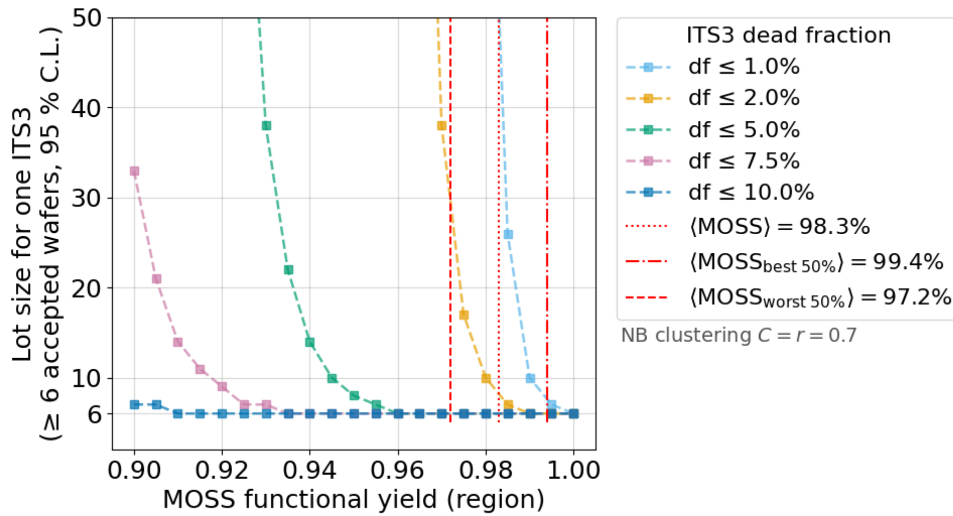


Figure B.2: ITS3 yield extrapolation and required wafer lot size considering different accepted ITS3 dead tile fractions using the Negative Binomial model extension with a clustering factor of $C = 0.7$.

Abbreviations

ADC . . . . Analog-to-Digital Converter
ALICE . . . . . . A Large Ion Collider Experiment
ALPIDE . . . . . ALICE PIxel DEtector (chip)
ANN . . . . Artificial Neural Network
API . . . . Application Programming Interface
BN . . . . Batch Normalisation
CB . . . . Conduction Band
CMOS . . . . . . Complementary Metal Oxide Semiconductor
CMP . . . . Chemical Mechanical Polishing
CTE . . . . Coefficient of Thermal Expansion
CVD . . . . Chemical Vapour Deposition
DAC . . . . Digital-to-Analog Converter
DCA . . . . Distance of Closest Approach
DNN . . . . Deep Neural Network
EDS . . . . Energy Dispersive X-ray spectroscopy
ER . . . . Engineering Run
ESD . . . . Electrostatic Discharge
FHR . . . . Fake-Hit Rate
FIB . . . . Focussed Ion Beam
FPC . . . . Flexible Printed Circuit
FPGA . . . . Field Programmable Gate Array
HU . . . . Half Unit
IB . . . . Inner Barrel
IMD . . . . Inter-Metal Dielectric
I/O . . . . Input/Output
IP . . . . Interaction Point
ITS . . . . Inner Tracking System
LDO . . . . Low-Dropout regulator
LEC . . . . Left Endcap
LHC . . . . Large Hadron Collider
LS . . . . Long Shutdown
M# . . . . Metal layer in metal stack (numbered)
MC . . . . Monte Carlo
MAPS . . . . Monolithic Active Pixel Sensor

MIP . . . . Minimum Ionising Particle
MLP . . . . Multi-Layer Perceptron
MOSS . . . . . . MOnolithic Stitched Sensor
MPV . . . . Most Probable Value
MS . . . . Multiple Scattering
NTC . . . . Negative Temperature Coefficient (thermistor)
OB . . . . Outer Barrel
PCB . . . . Printed Circuit Board
PCM . . . . Process Control Monitoring
PDF . . . . Probability Density Function
PID . . . . Particle Identification
PSUB . . . . P-Substrate (chip substrate)
PV . . . . Primary Vertex
PVD . . . . Physical Vapour Deposition
QCD . . . . Quantum Chromodynamics
QGP . . . . Quark-Gluon Plasma
REC . . . . Right Endcap
RMS . . . . Root Mean Square
ROI . . . . Region Of Interest
RSU . . . . Repeated Sensor Unit
S/B . . . . Signal-over-Background
SEM . . . . Scanning Electron Microscopy
SM . . . . Standard Model
SMU . . . . Source Measure Unit
SV . . . . Secondary Vertex
TPSCo . . . . . . Tower Partners Semiconductor Company
TPC . . . . Time Projection Chamber
UV . . . . Ultraviolet
VB . . . . Valence Band

References

[1] G. H. Eberwein and S. D'Auria. *Interlock requirements for site qualification*. Tech. rep. AT2-IP-QA-0031. Internal document. ATLAS Collaboration, CERN, Sept. 2022.

[2] R. Bates et al. *Report on the bumps stress in the ITk pixel detector*. Tech. rep. AT2-IP-EP-0011. Internal document. ATLAS Collaboration, CERN, Feb. 2022.

[3] G. Aglieri Rinella et al. "Characterization of analogue Monolithic Active Pixel Sensor test structures implemented in a 65 nm CMOS imaging process". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1069 (2024). DOI: 10.1016/j.nima.2024.169896.

[4] W. Snoeys et al. "Optimization of a 65 nm CMOS imaging process for monolithic CMOS sensors for high energy physics". In: *PoS* Pixel2022 (2023). DOI: 10.22323/1.420.0083.

[5] ALICE ITS3 WP3. *MOSS testing: software suite for MOSS chip characterisation*. https://gitlab.cern.ch/alice-its3-wp3/moss-testing. GitLab repository, CERN, accessed 27 Jul 2025. 2025.

[6] ALICE Collaboration. *Technical Design report for the ALICE Inner Tracking System 3 - ITS3 ; A bent wafer-scale monolithic pixel detector*. Tech. rep. Co-project Manager: Magnus Mager. Geneva: CERN, 2024. URL: https://cds.cern.ch/record/2890181.

[7] ALICE Collaboration. "Characterisation of the first wafer-scale prototype for the ALICE ITS3 upgrade: the monolithic stitched sensor (MOSS)". In submission to Nucl. Instrum. Methods Phys. Res. A. 2025.

[8] G. H. Eberwein. "A Novel Approach for Fault Detection and Failure Analysis of CMOS Copper Metal Stacks". In submission to IEEE TNS. 2025.

[9] M. E. Peskin and D. V. Schroeder. *An introduction to quantum field theory*. Boulder, CO: Westview, 1995. URL: https://cds.cern.ch/record/257493.

[10] M. K. Gaillard, P. D. Grannis, and F. J. Sciulli. "The standard model of particle physics". In: *Rev. Mod. Phys.* 71 (2 Mar. 1999). DOI: 10.1103/RevModPhys.71.S96.

[11] M. Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. ISBN: 9781107034266.

[12] 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). DOI: 10.1016/j.physletb.2012.08.020.

[13] 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). DOI: 10.1016/j.physletb.2012.08.021.

[14] Z. Citron et al. *Report from Working Group 5: Future physics opportunities for high-density QCD at the LHC with heavy-ion and proton beams*. Report on the Physics at the HL–LHC, and Perspectives for the HE–LHC CERN–LPCC–2018–07. Geneva, Switzerland: CERN, Dec. 2019, pp. 1159–1410. DOI: 10.23731/CYRM-2019-007.1159.

[15] ALICE Collaboration. *Letter of intent for ALICE 3: A next-generation heavy-ion experiment at the LHC*. 2022. arXiv: 2211.02491 [physics.ins-det].

[16] ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008). DOI: 10.1088/1748-0221/3/08/S08002.

[17] ALICE Collaboration. "The ALICE experiment: a journey through QCD". In: *Eur. Phys. J. C* 84 (2024). DOI: 10.1140/epjc/s10052-024-12935-y.

[18] I. Shipsey. "Where do we go from here? Vision & Outlook". In: *PoS* ICHEP2016 (2017). DOI: 10.22323/1.282.0037.

[19] B. Abelev et al. *Technical Design Report for the Upgrade of the ALICE Inner Tracking System*. Tech. rep. 2014. DOI: 10.1088/0954-3899/41/8/087002.

[20] F. Reidt. "Upgrade of the ALICE ITS detector". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1032 (2022). DOI: 10.1016/j.nima.2022.166632.

[21] L. Musa. *Letter of Intent for an ALICE ITS Upgrade in LS3*. Tech. rep. Geneva: CERN, 2019. DOI: 10.17181/CERN-LHCC-2019-018.

[22] P. V. Leitao et al. "Development of a Stitched Monolithic Pixel Sensor prototype (MOSS chip) towards the ITS3 upgrade of the ALICE Inner Tracking system". In: *Journal of Instrumentation* 18.01 (Jan. 2023). DOI: 10.1088/1748-0221/18/01/C01044.

[23] L. Gonella. "Development of a Silicon Vertex and Tracking Detector for the Electron-Ion Collider". In: *PoS* VERTEX2023 (2024). DOI: 10.22323/1.448.0038.

[24] M. Benedikt. *Future Circular Collider Feasibility Study Report Volume 1 : Physics and Experiments*. Version 2.0. Mar. 2025. DOI: 10.17181/CERN.9DKX.TDH9.

[25] E. Lopienska. "The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022". In: (2022). General Photo. URL: https://cds.cern.ch/record/2800984.

[26] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008). DOI: 10.1088/1748-0221/3/08/S08001.

[27] L. Arnaudon et al. *Linac4 Technical Design Report*. Tech. rep. revised version submitted on 2006-12-14 09:00:40. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/1004186.

[28] CERN. *The third run of the Large Hadron Collider has successfully started*. https://home.cern/news/news/cern/third-run-large-hadron-collider-has-successfully-started. Stable pp collisions at $\sqrt{s} = 13.6$ TeV, 5 July 2022. Accessed: 20 June 2025. 2022.

[29] R. Bruce et al. "First results of running the LHC with lead ions at a beam energy of 6.8 Z TeV". In: *JACoW IPAC* 2023 (2023). DOI: 10.18429/JACoW-IPAC2023-MOPL021.

[30] G. Aad et al. “The ATLAS Experiment at the CERN Large Hadron Collider”. In: *JINST* 3 (2008). Also published by CERN Geneva in 2010. DOI: 10.1088/1748-0221/3/08/S08003.

[31] S. Chatrchyan et al. “The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment”. In: *JINST* 3 (2008). Also published by CERN Geneva in 2010. DOI: 10.1088/1748-0221/3/08/S08004.

[32] A. A. Alves et al. “The LHCb Detector at the LHC”. In: *JINST* 3 (2008). Also published by CERN Geneva in 2010. DOI: 10.1088/1748-0221/3/08/S08005.

[33] S. Navas et al. “Review of particle physics”. In: *Phys. Rev. D* 110.3 (2024). DOI: 10.1103/PhysRevD.110.030001.

[34] CERN Beam Performance Tracking Team. *LHC Statistics*. https://bpt.web.cern.ch/lhc/statistics/2024/. Beam Performance Tracking in the CERN accelerator complex. Accessed 21 June 2025. 2025.

[35] ATLAS Collaboration. *Total Integrated Luminosity in Run 3 (13.6 TeV pp data only): Interactions per Crossing 2022–2024*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun3. ATLAS Public Luminosity Results Run 3. Accessed 21 June 2025. 2025.

[36] CERN LHCC. *161st LHCC Meeting – OPEN Session (3–6 March 2025)*. https://indico.cern.ch/event/1509225/. CERN; included updates on LHC Run3 performance. CERN, Mar. 2025.

[37] LHC Programme Coordination (LPC), CERN. *Heavy-Ion and pp reference run configuration for Run 3 (2024)*. https://lpc.web.cern.ch/Run3/HIConfiguration2024.html. Details LHC beam energy, bunch structure, luminosity levelling, and filling schemes for Pb–Pb and pp runs. Accessed 21 June 2025. 2024.

[38] C. Pralavorio. “For one day only LHC collides xenon beams ”. In: (2017). URL: https://cds.cern.ch/record/2291272.

[39] D. Dobrigkeit Chinellato, S. Pisano, and T. Nayak. “2025-Oxygen-Run”. https://alice-collaboration.web.cern.ch/2025-LHC-Oxygen2025. Accessed 08 July 2025. 2025. URL: https://cds.cern.ch/record/2937161.

[40] G. Barr et al. *Particle Physics in the LHC Era*. Oxford University Press, Jan. 2016. ISBN: 9780198748557. DOI: 10.1093/acprof:oso/9780198748557.001.0001.

[41] J. Campbell, J. Huston, and F. Krauss. *The Black Book of Quantum Chromodynamics: A Primer for the LHC Era*. Oxford University Press, Dec. 2017. ISBN: 9780199652747. DOI: 10.1093/oso/9780199652747.001.0001.

[42] P. Braun-Munzinger and J. Stachel. “The quest for the quark–gluon plasma”. In: *Nature* 448.7151 (2007). DOI: 10.1038/nature06080.

[43] H. G. Ritter. “Prospects for an energy scan program at RHIC”. In: *PoS* CPOD2006 (2007). DOI: 10.22323/1.029.0015.

[44] H.-T. Ding, F. Karsch, and S. Mukherjee. “Thermodynamics of strong-interaction matter from lattice QCD”. In: *International Journal of Modern Physics E* 24.10 (2015). Cited by: 291; All Open Access, Green Open Access. DOI: 10.1142/S0218301315300076.

[45] P. F. Kolb and U. Heinz. *Hydrodynamic description of ultrarelativistic heavy-ion collisions*. 2003. arXiv: nucl-th/0305084 [nucl-th].

[46] P. Braun-Munzinger, K. Redlich, and J. Stachel. "PARTICLE PRODUCTION IN HEAVY ION COLLISIONS". In: *Quark–Gluon Plasma 3*. WORLD SCIENTIFIC, Jan. 2004, pp. 491–599. DOI: 10.1142/9789812795533_0008.

[47] J. W. Harris and B. Müller. ""QGP Signatures" Revisited". In: *Eur. Phys. J. C* 84.3 (2024). DOI: 10.1140/epjc/s10052-024-12533-y.

[48] S. Acharya et al. "Measurement of dielectron production in central Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV". In: *Phys. Rev. C* 99 (2019). 34 pages, 14 captioned figures, 3 tables, authors from page 29, published, figures at http://alice-publications.web.cern.ch/node/4474. DOI: 10.1103/PhysRevC.99.024002.

[49] A. Andronic et al. "Heavy-flavour and quarkonium production in the LHC era: from proton–proton to heavy-ion collisions". In: *The European Physical Journal C* 76.3 (2016). DOI: 10.1140/epjc/s10052-015-3819-5.

[50] "Upgrade of the ALICE Inner Tracking System during LS3: study of physics performance". In: (2023). URL: https://cds.cern.ch/record/2868015.

[51] S. Acharya et al. "First Measurement of $A = 4$ Hypernuclei and Antihypernuclei at the LHC". In: *Phys. Rev. Lett.* 134 (16 Apr. 2025). DOI: 10.1103/PhysRevLett.134.162301.

[52] S. Acharya et al. "ALICE upgrades during the LHC Long Shutdown 2". In: *Journal of Instrumentation* 19.05 (May 2024). DOI: 10.1088/1748-0221/19/05/P05062.

[53] ALICE TPC Collaboration. "The upgrade of the ALICE TPC with GEMs and continuous readout". In: *Journal of Instrumentation* 16.03 (Mar. 2021). DOI: 10.1088/1748-0221/16/03/P03022.

[54] ALICE Collaboration. "Technical Design Report for the Upgrade of the ALICE Inner Tracking System". In: *Journal of Physics G: Nuclear and Particle Physics* 41.8 (July 2014). DOI: 10.1088/0954-3899/41/8/087002.

[55] M. Mager. "ALPIDE, the Monolithic Active Pixel Sensor for the ALICE ITS upgrade". In: *Nucl. Instrum. Methods Phys. Res., A* 824 (2016). DOI: 10.1016/j.nima.2015.09.057.

[56] CERN. *CERN Courier Volume 61, Number 4, July/August 2021*. 2021. URL: https://cds.cern.ch/record/2773907.

[57] ALICE Collaboration. "Performance of the ALICE experiment at the CERN LHC". In: *International Journal of Modern Physics A* 29.24 (2014). DOI: 10.1142/S0217751X14300440.

[58] C.-Y. Wong. *Introduction to high-energy heavy-ion collisions*. Singapore: World Scientific, 1991. ISBN: 9810202636.

[59] H. Kolanoski and N. Wermes. *Particle Detectors: Fundamentals and Applications*. Oxford University Press, June 2020. ISBN: 9780198858362. DOI: 10.1093/oso/9780198858362.001.0001.

[60] I. Neutelings. *CMS coordinate system and Pseudorapidity – TikZ.net*. https://tikz.net/axis3d_cms/. Illustration from TikZ.net; retrieved June 17, 2025. 2021.

[61] H. A. Bethe. “Molière’s Theory of Multiple Scattering”. In: *Phys. Rev.* 89 (6 Mar. 1953). DOI: 10.1103/PhysRev.89.1256.

[62] N. Valle. “Performance of the ALICE Inner Tracking System 2”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1079 (2025). DOI: 10.1016/j.nima.2025.170596.

[63] H. Spieler. *Semiconductor detector systems*. Series on semiconductor science and technology. Oxford: Oxford Univ. Press, 2005. DOI: 10.1093/acprof:oso/9780198527848.001.0001.

[64] R. Frühwirth and A. Strandlie. “Pattern Recognition and Reconstruction”. In: *Detectors for Particles and Radiation. Part1: Principles and Methods*. Ed. by C. Fabjan and H. Schopper. Vol. 21B1. Landolt–Börnstein, Group I: Elementary Particles, Nuclei and Atoms. Springer Berlin Heidelberg, 2011, pp. 352–387. DOI: 10.1007/978-3-642-03606-4.

[65] Z. Drasal and W. Riegler. “An extension of the Gluckstern formulae for multiple scattering: Analytic expressions for track parameter resolution using optimum weights”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 910 (2018). DOI: 10.1016/j.nima.2018.08.078.

[66] M. Mager. “The future of bent MAPS, full-wafer (stitched) design: Status and challenges”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1064 (2024). DOI: 10.1016/j.nima.2024.169447.

[67] ALICE ITS project. “First demonstration of in-beam performance of bent Monolithic Active Pixel Sensors”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1028 (2022). DOI: 10.1016/j.nima.2021.166280.

[68] A. Kluge. “ALICE - ITS3 — A bent, wafer-scale CMOS detector”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1041 (2022). DOI: 10.1016/j.nima.2022.167315.

[69] A. Amatriain et al. “Development of an air cooling system with low material budget for high-energy physics applications”. In: *Applied Thermal Engineering* 236 (2024). DOI: 10.1016/j.applthermaleng.2023.121699.

[70] A. Amatriain et al. “Aeroelastic analysis using confocal sensors: experimental study and numerical validation with application to a future particle detector”. In: *Aerospace Science and Technology* 151 (2024). DOI: 10.1016/j.ast.2024.109313.

[71] C. Gargiulo, A. Junique, and G. H. Eberwein. “ALICE ITS3 prototype setup”. General Photo. 2024. URL: https://cds.cern.ch/record/2895742.

[72] P. Buncic, M. Krzewicki, and P. Vande Vyvre. *Technical Design Report for the Upgrade of the Online-Offline Computing System*. Tech. rep. 2015. URL: https://cds.cern.ch/record/2011297.

[73] Ananya et al. “O2: A novel combined online and offline computing system for the ALICE Experiment after 2018”. In: *Journal of Physics: Conference Series* 513.1 (June 2014). DOI: 10.1088/1742-6596/513/1/012037.

[74] C. Bierlich et al. *A comprehensive guide to the physics and usage of PYTHIA 8.3*. 2022. arXiv: 2203.11601 [hep-ph].

[75] X.-N. Wang and M. Gyulassy. "hijing: A Monte Carlo model for multiple jet production in pp, pA, and AA collisions". In: *Phys. Rev. D* 44 (11 Dec. 1991). DOI: 10.1103/PhysRevD.44.3501.

[76] 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 (2003). DOI: 10.1016/S0168-9002(03)01368-8.

[77] M. Puccio. "Tracking in high-multiplicity events". In: *PoS* Vertex 2016 (2017). DOI: 10.22323/1.287.0043.

[78] S. Radhakrishnan et al. "Measurements of open charm production in Au+Au collisions at psNN = 200 GeV with the STAR experiment at RHIC". In: Brookhaven National Lab. (BNL), Upton, NY (United States). May 2018. DOI: 10.1016/j.nuclphysa.2018.10.050.

[79] G. M. Innocenti. "Latest results on Λ_c^+ and D production in pp and Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$TeV with ALICE at the LHC". In: *Nucl. Phys. A* 1005 (2021). DOI: 10.1016/j.nuclphysa.2020.122002.

[80] S. Acharya et al. "Constraining hadronization mechanisms with Λc+/D0 production ratios in Pb–Pb collisions at sNN=5.02 TeV". In: *Physics Letters B* 839 (2023). DOI: 10.1016/j.physletb.2023.137796.

[81] C. Zampolli. "Latest ALICE results on D_s and Λ_c production in Pb–Pb collisions at 5.02 TeV". https://indico.cern.ch/event/755366/contributions/3396487/. Presented at Strangeness in Quark Matter (SQM) 2019, June 2019. 2019.

[82] S. Acharya et al. "Λ_c^+ Production and Baryon-to-Meson Ratios in pp and p-Pb Collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV at the LHC". In: *Phys. Rev. Lett.* 127 (20 Nov. 2021). DOI: 10.1103/PhysRevLett.127.202301.

[83] P. D. Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). DOI: 10.1093/ptep/ptaa104.

[84] R. Aaij et al. "Amplitude analysis of the $\Lambda_c^+ \to pK^-\pi^+$ decay and Λ_c^+ baryon polarization measurement in semileptonic beauty hadron decays". In: *Phys. Rev. D* 108 (1 July 2023). DOI: 10.1103/PhysRevD.108.012023.

[85] H. Bichsel. "Straggling in thin silicon detectors". In: *Rev. Mod. Phys.* 60 (3 July 1988). DOI: 10.1103/RevModPhys.60.663.

[86] F. Wang et al. "The impact of incorporating shell-corrections to energy loss in silicon". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 899 (2018). DOI: 10.1016/j.nima.2018.04.061.

[87] W. R. Leo. *Techniques for nuclear and particle physics experiments: a how-to approach; 2nd ed.* Berlin: Springer, 1994. DOI: 10.1007/978-3-642-57920-2.

[88] L. Rossi et al. *Pixel detectors: from fundamentals to applications*. Particle acceleration and detection. Berlin: Springer, 2006. DOI: 10.1007/3-540-28333-1.

[89] M. Berger et al. *XCOM: Photon Cross Section Database (version 1.2)*. en. 1999-01-01 1999.

[90] J. R. Hook and H. E. Hall. *Solid state physics*. eng. 2nd ed., reprinted with corrections. Manchester physics series. Chichester: Wiley, 2000 - 1991. ISBN: 0471928046.

[91] M. A. Green. "Intrinsic concentration, effective densities of states, and effective mass in silicon". In: *Journal of Applied Physics* 67.6 (Mar. 1990). DOI: 10.1063/1.345414.

[92] D. A. Neamen. *Semiconductor physics and devices*. eng. 4th edition. New York: McGraw-Hill Higher Education, 2012. ISBN: 9780073529585.

[93] C. Kittel and P. McEuen. *Introduction to solid state physics*. eng. 8th ed. New York ; Wiley, 2005. ISBN: 047141526X.

[94] D. Groom. *Temperature dependence of mean number of e–h pairs per eV of X–ray energy deposit*. Technical Note. Accessed 26 Jun 2025. Lawrence Berkeley National Laboratory, Dec. 2004. URL: https://www-ccd.lbl.gov/w_Si.pdf.

[95] G. F. Knoll. *Radiation detection and measurement; 4th ed.* New York, NY: Wiley, 2010. URL: https://cds.cern.ch/record/1300754.

[96] M. Moll. "Displacement Damage in Silicon Detectors for High Energy Physics". In: *IEEE Transactions on Nuclear Science* 65.8 (2018). DOI: 10.1109/TNS.2018.2819506.

[97] C. Canali et al. "Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature". In: *IEEE Transactions on Electron Devices* 22.11 (1975). DOI: 10.1109/T-ED.1975.18267.

[98] C. Kittel and P. McEuen. *Introduction to solid state physics*. eng. 8th ed. New York ; Wiley, 2005. ISBN: 047141526X.

[99] J. Fossum and D. Lee. "A physical model for the dependence of carrier lifetime on doping density in nondegenerate silicon". In: *Solid-State Electronics* 25.8 (1982). DOI: 10.1016/0038-1101(82)90203-9.

[100] W. Shockley and W. T. Read. "Statistics of the Recombinations of Holes and Electrons". In: *Phys. Rev.* 87 (5 Sept. 1952). DOI: 10.1103/PhysRev.87.835.

[101] G. Aglieri Rinella. "The ALPIDE pixel sensor chip for the upgrade of the ALICE Inner Tracking System". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 845 (2017). Proceedings of the Vienna Conference on Instrumentation 2016. DOI: 10.1016/j.nima.2016.05.016.

[102] G. Aglieri Rinella et al. "Charge collection properties of TowerJazz 180 nm CMOS Pixel Sensors in dependence of pixel geometries and bias parameters, studied using a dedicated test-vehicle: the Investigator chip." In: *Nucl. Instrum. Methods Phys. Res., A* 988 (2021). DOI: 10.1016/j.nima.2020.164859.

[103] W. Snoeys. "CMOS monolithic active pixel sensors for high energy physics". In: *Nucl. Instrum. Methods Phys. Res., A* 765 (2014). DOI: 10.1016/j.nima.2014.07.017.

[104] W. Snoeys et al. "A process modification for CMOS monolithic active pixel sensors for enhanced depletion, timing performance and radiation tolerance". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 871 (2017). DOI: 10.1016/j.nima.2017.07.046.

[105] M. Munker et al. “Simulations of CMOS pixel sensors with a small collection electrode, improved for a faster charge collection and increased radiation tolerance”. In: *Journal of Instrumentation* 14.05 (May 2019). DOI: 10.1088/1748-0221/14/05/C05013.

[106] F. Piro et al. “A Compact Front-End Circuit for a Monolithic Sensor in a 65-nm CMOS Imaging Technology”. In: *IEEE Transactions on Nuclear Science* 70.9 (2023). DOI: 10.1109/TNS.2023.3299333.

[107] G. A. Rinella et al. “Digital Pixel Test Structures implemented in a 65 nm CMOS process”. In: *Nucl. Instrum. Methods Phys. Res., A* 1056 (2023). DOI: 10.1016/j.nima.2023.168589.

[108] G. A. Rinella et al. “Time performance of Analog Pixel Test Structures with in-chip operational amplifier implemented in 65 nm CMOS imaging process”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1070 (2025). DOI: 10.1016/j.nima.2024.170034.

[109] R. Baker. *CMOS: Circuit Design, Layout, and Simulation, Third Edition*. Vol. 18. John Wiley I& Sons, Sept. 2010. ISBN: 978-0470881323. DOI: 10.1002/9780470891179.

[110] G. May and C. Spanos. *Fundamentals of Semiconductor Manufacturing and Process Control*. May 2006. ISBN: 9780471784067. DOI: 10.1002/0471790281.index.

[111] M. I. Bâzu and T. Băjenescu. *Failure Analysis: A Practical Guide for Manufacturers of Electronic Components and Systems*. Hoboken, NJ: John Wiley & Sons, 2011. DOI: 10.1002/9781119990093.

[112] H. H. Radamson. *CMOS past, present and future*. eng. Woodhead Publishing Series in Electronic and Optical Materials. Duxford, United Kingdom: Woodhead Publishing, An imprint of Elsevier, 2018. ISBN: 9780081021408.

[113] J. Segura and C. F. Hawkins. *CMOS Electronics: How It Works, How It Fails*. Hoboken, NJ: John Wiley & Sons, 2004, p. 366. ISBN: 978-0-471-47669-6. DOI: 10.1002/0471728527.

[114] H. J. Veendrick. *Nanometer CMOS ICs: From Basics to ASICs*. 2nd ed. Second edition; 648 pp. including preliminary matter. Cham, Switzerland: Springer, 2017, pp. xxxvii + 611. ISBN: 978-3-319-47595-0. DOI: 10.1007/978-3-319-47597-4.

[115] J. D. Plummer and P. B. Griffin. *Integrated Circuit Fabrication: Science and Technology*. 1st ed. Print edition. Cambridge, England: Cambridge University Press, 2024. ISBN: 9781009157852.

[116] H. Geng, ed. *Semiconductor Manufacturing Handbook*. 2nd. New York: McGraw-Hill Education, 2018. ISBN: 9781259587696.

[117] S. Natarajan et al. “A 32nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171μm2 SRAM cell size in a 291Mb array”. In: *2008 IEEE International Electron Devices Meeting*. 2008, pp. 1–3. DOI: 10.1109/IEDM.2008.4796777.

[118] D. Zhao and X. Lu. “Chemical mechanical polishing: Theory and experiment”. In: *Friction* 1.4 (2013). DOI: 10.1007/s40544-013-0035-x.

[119] J. Gambino. “Chapter 6 - Process Technology for Copper Interconnects”. In: *Handbook of Thin Film Deposition (Fourth Edition)*. Ed. by K. Seshan and D. Schepis. Fourth Edition. William Andrew Publishing, 2018, pp. 147–194. DOI: 10.1016/B978-0-12-812311-9.00006-2.

[120] E. Ogawa et al. “Stress-induced voiding under vias connected to wide Cu metal leads”. In: *2002 IEEE International Reliability Physics Symposium. Proceedings. 40th Annual (Cat. No.02CH37320)*. 2002, pp. 312–321. DOI: 10.1109/RELPHY.2002.996654.

[121] Y.-H. C. et al. *Two-Stage Cu Anneal to Improve Cu Damascene Process*. US 6,391,777 B1. May 2002.

[122] S. Chambers and others. *Methods and Devices for the Suppression of Copper Hillock Formation*. US 6,846,752 B2. Jan. 2005.

[123] J. W. et al. *Low temperature method for minimizing copper hillock defects*. US 2006/0252258 A1. Nov. 2006.

[124] G. Aglieri Rinella. “Developments of stitched monolithic pixel sensors towards the ALICE ITS3”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1049 (2023). DOI: 10.1016/j.nima.2023.168018.

[125] S. E. Bohndiek et al. “Characterization and Testing of LAS: A Prototype ‘Large Area Sensor’ With Performance Characteristics Suitable for Medical Imaging Applications”. In: *IEEE Transactions on Nuclear Science* 56.5 (2009). DOI: 10.1109/TNS.2009.2029575.

[126] R. Turchetta, N. Guerrini, and I. Sedgwick. “Large area CMOS image sensors”. In: *Journal of Instrumentation* 6.01 (Jan. 2011). DOI: 10.1088/1748-0221/6/01/C01099.

[127] D. Durini et al. “Large full-well capacity stitched CMOS image sensor for high temperature applications”. In: *2010 Proceedings of ESSCIRC*. 2010, pp. 130–133. DOI: 10.1109/ESSCIRC.2010.5619823.

[128] M. Esposito et al. “Performance of a novel wafer scale CMOS active pixel sensor for bio-medical imaging”. In: *Physics in Medicine I& Biology* 59.13 (June 2014). DOI: 10.1088/0031-9155/59/13/3533.

[129] C. Systems. *WSE-3 Datasheet*. Tech. rep. WSE-3. Cerebras Systems, 2024. URL: https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/Datasheets/WSE-3%20Datasheet.pdf.

[130] J. J. Hoogstra, R. B. Anderson, and A. D. Graham. “Stitching Design Rules for Forming Interconnect Layers”. US6225013B1. May 2001. URL: https://patentimages.storage.googleapis.com/86/09/15/25326cd6779505/US6225013B1.pdf.

[131] J. Zhu et al. “Systematic experimental study on stitching techniques of CMOS image sensors”. In: *IEICE Electronics Express* 13.15 (2016). DOI: 10.1587/elex.13.20160441.

[132] P. Vicente Leitao. “Development of the MOSAIX chip for the ALICE ITS3 upgrade”. In: *JINST* 20.06 (2025). DOI: 10.1088/1748-0221/20/06/C06001.

[133] G. Aglieri Rinella et al. “Power distribution over the wafer-scale monolithic pixel detector — MOSAIX for ALICE ITS3”. In: *Journal of Instrumentation* 20.02 (Feb. 2025). DOI: 10.1088/1748-0221/20/02/C02015.

[134] P. Dorosz, on behalf of the MOSIAX design team, and the ALICE collaboration. “Data transmission architecture of the ALICE ITS3 stitched sensor prototype MOSAIX”. In: *Journal of Instrumentation* 19.04 (Apr. 2024). DOI: 10.1088/1748-0221/19/04/C04050.

[135] V. Sarritzu. “ITS3: the ALICE Inner Tracking System upgrade”. In: *PoS* TIPP2023 (2025). DOI: 10.22323/1.468.0126.

[136] DISCO Corporation. *Corporate Outline*. https://www.disco.co.jp/eg/corporate/outline/index.html. Accessed 3 July 2025.

[137] Semiconductor Equipment Corporation. *Die Ejector Model 4800*. Die Ejector. Semiconductor Equipment Corporation. 2025. URL: https://www.semicorp.com/wp-content/uploads/2020/06/SEC-Model-4800-060220.pdf.

[138] J. Schmalz GmbH. *Bellows suction cup (round) FSGA 2 NBR-55 M3-AG*. https://www.schmalz.com/10.01.06.04521. Part no. 10.01.06.04521, accessed 3 July 2025.

[139] Adafruit Industries. *Air Pump and Vacuum DC Motor - 4.5V and 1.8 LPM (ZR320-02PM)*. Product ID: 4700. Micro air pump suitable for inflatables and air-powered projects. 2020. URL: https://www.adafruit.com/product/4700.

[140] Arduino. *Arduino Leonardo*. Microcontroller board based on the ATmega32u4. Arduino. 2012. URL: https://docs.arduino.cc/hardware/leonardo/.

[141] Adafruit Industries. *Adafruit Motor/Stepper/Servo Shield for Arduino v2 Kit - v2.3*. Product ID: 1438. Shield for controlling DC motors, stepper motors, and servos via Arduino. Adafruit Industries. 2013. URL: https://www.adafruit.com/product/1438.

[142] Analog Devices. *DC2026C (Linduino One) Evaluation Board*. Arduino-compatible development platform with isolated USB interface. Analog Devices. 2016. URL: https://www.analog.com/en/resources/evaluation-hardware-and-software/evaluation-boards-kits/dc2026c.html.

[143] Delta Line. *42SH47-4AM-PZK2869KZCZK138 Stepper Motor*. Size 42 mm (NEMA 17), 0.9° step angle, 1.68 A/phase, 0.44 Nm holding torque. Delta Line. 2020. URL: https://en.delta-line.com/a.pag/42sh47-4am-pzk2869kzczk138.html.

[144] STMicroelectronics. *STSPIN820: Stepper Motor Driver Datasheet*. Datasheet. Document no. DS12345 Rev 3. STMicroelectronics, 2018. URL: https://www.st.com/resource/en/datasheet/stspin820.pdf.

[145] Zaber Technologies Inc. *X-RSW60A Motorized Rotary Stage Datasheet*. 60 mm stage top, 225 N-cm torque, 0.08° accuracy, integrated controller. Zaber Technologies Inc. 2023. URL: https://www.zaber.com/api/assets/X-RSW60A-Datasheet.pdf.

[146] Norland Products Inc. *Norland Optical Adhesive 68: Technical Data Sheet*. UV-curable adhesive designed for bonding glass to plastics and metals. Norland Products Inc. 2025. URL: https://norlandproducts.com/wp-content/uploads/2025/02/NOA-68-TDS.pdf.

[147] Keysight Technologies. *B2901B Precision Source/Measure Unit: Data Sheet*. Data Sheet 3120-1466EN. Published August 2, 2022. Accessed 3 July 2025. USA: Keysight Technologies, Aug. 2022.

[148] Keysight Technologies. *DAQ970A/DAQ973A Data Acquisition System: Technical Overview*. Technical Overview 5992-3168EN. Published January 20, 2025. Accessed 3 July 2025. Configuration used in this work: 3x DAQM901A 20(22)-channel multiplexer module. USA: Keysight Technologies, Jan. 2025.

[149] Rohde & Schwarz GmbH & Co. KG. *R&S® HMP4040 Power Supply Series: Data Sheet*. Data Sheet. Version 02.01, accessed 3 July 2025. Configuration used in this work: model HMP4040 (four-channel version). Munich, Germany: Rohde & Schwarz, June 2022.

[150] Keithley Instruments. *DMM7510 7½-Digit Graphical Sampling Multimeter: Data Sheet*. Data Sheet 1KW-60022-3. Published April 14, 2023. Accessed 3 July 2025. Cleveland, OH, USA: Keithley Instruments (a Tektronix Company), Apr. 2023.

[151] FLIR Systems, Inc. *A655sc High-Resolution LWIR Science-Grade Infrared Camera: Data Sheet*. Data Sheet 3445 Rev. 4/13. Accessed 3 July 2025. Wilsonville, OR, USA: FLIR Systems, Inc., Apr. 2013.

[152] FLIR Systems, Inc. *Close-up IR Lens, 2.9× (50 μm) with case*. Data Sheet T198059. Published June 22 2014 (document ID T198059). Accessed 3 July 2025. Wilsonville, OR, USA: FLIR Systems, Inc., June 2014.

[153] G. Bradski. “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (2000).

[154] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. 4th ed. Pearson, 2018. ISBN: 9780133356724.

[155] A. Buades, B. Coll, and J.-M. Morel. “Non-Local Means Denoising”. In: *Image Processing On Line* 1 (2011). `https://doi.org/10.5201/ipol.2011.bcm_nlm`.

[156] M. Köfferlein. *KLayout*. `https://www.klayout.de`. Version 0.28.17. 2024.

[157] Enclustra GmbH. *Mercury+ PE1-200/300/400 Base Board User Manual*. Version V06. Accessed 3 July 2025. Zürich, Switzerland. URL: `https://www.enclustra.com/en/products/base-boards/mercury-pe1-200-300-400/`.

[158] Enclustra GmbH. *Mercury+ AA1 Intel Arria 10 System-on-Chip Module User Manual*. Version V09. Accessed 3 July 2025. Zürich, Switzerland. URL: `https://www.enclustra.com/en/products/system-on-chip-modules/mercury-aa1/`.

[159] TDK Corporation. *NTC Thermistors, NTCG Series*. Accessed 3 July 2025. Tokyo, Japan, 2023. URL: `https://product.tdk.com/system/files/dam/doc/product/sensor/ntc/chip-ntc-thermistor/catalog/tpd_automotive_ntc-thermistor_ntcg_en.pdf`.

[160] S. Bugiel. Private communication. Feb. 2024.

[161] R. F. Tate. “Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation”. In: *The Annals of Mathematical Statistics* 25.3 (1954). URL: `http://www.jstor.org/stable/2236844` (visited on 03/16/2025).

[162] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947). URL: http://www.jstor.org/stable/2236101 (visited on 03/16/2025).

[163] A. A. Hagberg, D. A. Schult, and P. J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.

[164] J. Moyal. "Theory of ionization fluctuations". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 46.374 (1955). DOI: 10.1080/14786440308521076.

[165] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020). DOI: 10.1038/s41592-019-0686-2.

[166] R. Storn and K. Price. "Differential Evolution - A Simple and Efficient Heuristic for global Optimization over Continuous Spaces". In: *Journal of Global Optimization* 11.4 (1997). DOI: 10.1023/A:1008202821328.

[167] G. H. Eberwein. *Yield Characterisation and Failure Analysis of the Monolithic Stitched Sensor MOSS for ALICE ITS3*. Oct. 2024. URL: https://indico.cern.ch/event/1381495/contributions/5988500 (visited on 07/20/2025).

[168] Carl Zeiss Microscopy GmbH. *ZEISS Crossbeam FIB-SEM for High Throughput 3D Analysis and Sample Preparation*. https://www.zeiss.com/microscopy/en/products/sem-fib-sem/fib-sem/crossbeam.html. Accessed: 2025-07-08. 2025.

[169] F. James. *MINUIT: Function Minimization and Error Analysis*. Version 94.1. CERN Program Library Long Writeups D506. CERN. Geneva, Switzerland, Aug. 1998. URL: https://root.cern.ch/download/minuit.pdf.

[170] R. Brun and F. Rademakers. "ROOT — An object oriented data analysis framework". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.1 (1997). New Computing Techniques in Physics Research V. DOI: 10.1016/S0168-9002(97)00048-X.

[171] B. Yi, Y. Li, and T. C. M. Lee. "Multilayer Perceptrons: An Introduction". In: *Wiley StatsRef: Statistics Reference Online*. John Wiley I& Sons, Ltd, 2023, pp. 1–10. DOI: 10.1002/9781118445112.stat08394.

[172] A. Subasi. "Chapter 3 - Machine learning techniques". In: *Practical Machine Learning for Data Analysis Using Python*. Ed. by A. Subasi. Academic Press, 2020, pp. 91–202. DOI: 10.1016/B978-0-12-821379-7.00003-5.

[173] D. Sarkar, R. Bali, and T. Sharma. *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems*. 1st. USA: Apress, 2017. ISBN: 1484232062.

[174] C. Garbin, X. Zhu, and O. Marques. "Dropout vs. batch normalization: an empirical study of their impact to deep learning". In: *Multimedia Tools and Applications* 79.19 (2020). DOI: 10.1007/s11042-019-08453-9.

[175] T. Kim. *Generalizing MLPs With Dropouts, Batch Normalization, and Skip Connections*. 2021. arXiv: 2108.08186 [cs.LG].

[176] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 315–323. URL: `https://proceedings.mlr.press/v15/glorot11a.html`.

[177] F. E. Bock et al. "Hybrid Modelling by Machine Learning Corrections of Analytical Model Predictions towards High-Fidelity Simulation Solutions". In: *Materials* 14.8 (2021). DOI: `10.3390/ma14081883`.

[178] A. Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: `1912.01703 [cs.LG]`.

[179] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: `1412.6980 [cs.LG]`.

[180] M. Cacciari, M. Greco, and P. Nason. "The pT spectrum in heavy-flavour hadroproduction". In: *Journal of High Energy Physics* 1998.05 (June 1998). DOI: `10.1088/1126-6708/1998/05/007`.

[181] S. Beolé et al. "The MAPS foil". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1046 (2023). DOI: `10.1016/j.nima.2022.167673`.

[182] A. Mapelli, P. Petagna, and M. Vos. *Micro-channel cooling for collider experiments: review and recommandations*. Tech. rep. Geneva: CERN, 2020. URL: `https://cds.cern.ch/record/2712079`.

[183] O. C. Ibe. "1 - Basic Concepts in Probability". In: *Markov Processes for Stochastic Modeling (Second Edition)*. Ed. by O. C. Ibe. Second Edition. Oxford: Elsevier, 2013, pp. 1–27. DOI: `10.1016/B978-0-12-407795-9.00001-3`.

[184] J. M. Hilbe. *Negative Binomial Regression*. 2nd ed. Cambridge: Cambridge University Press, 2011. DOI: `10.1017/CBO9780511973420`.

Acknowledgements

This thesis would not exist without the guidance, support, teamwork, and friendship of many who made my DPhil experience at Oxford and CERN the formative and invaluable period it was.

First, I want to thank my supervisor, Ian Shipsey, for leaving a lasting impact on my professional and personal development in the best way possible. I am grateful for the support, guidance, and opportunities I received, the discussions on physics and life, and for always leaving me with an excitement to continue my work and push further – even when meeting at 8 pm on a Friday. The inspiration and aspiration I experienced were unmatched, and I endeavoured to finish the thesis to your standards.

My heartfelt thanks go out to Daniela Bortoletto for picking up my supervision through it all, proofreading my thesis in detail, and guiding and supporting me through this last important stretch of my DPhil journey to the final thesis submission, while discussing the last details at any hour and time.

Magnus Mager for welcoming and including me in the excellent ITS3 team, providing me with supervision and an environment to make an impact in the ITS3 project – allowing me to take on exciting responsibilities. Further, for enabling me to extend my stay at CERN, ensuring I keep my focus when needed, providing valuable feedback, and offering opportunities to present my work.

I want to thank Antoine Junique not only for introducing me to proper cleanroom etiquette, showing me how to handle silicon, and spending hours evaluating new procedures while finding solutions to connect to output pins when there aren't any left – but also for helping me with my stranded car in necessary French, and showing me what Mont Blanc looks like up close from a plane.

Thanks go out to the entire WP3/ITS3 team: Alex Kluge, Antonello Di Mauro, Hartmut Hillemanns, Paolo Martinengo, Valerio Sarritzu, Ola Groettvik, Miljenko Suljic, Felix Reidt, Francesca Carnesecchi, Ivan Ravasenga, and many more, for fruitful and often vivid discussions, relentless debugging in the lab, finding the boundaries of what is pedantic in coding, and skiing in various conditions.

The brilliant chip design team around Walter Snoeys and Gianluca Aglieri Rinella, for answering my questions on chip manufacturing, fabrication, and design, providing feedback on all our measurements and figuring out a way to publish what shall not be published.

From the materials department, my thanks go out to Alice Moros and Anite Perez, carefully preparing my samples, trying new imaging approaches, and together staring determinedly at cross-section images until finding the structures I was so keen to prove exist.

For the guidance on the simulation work, I want to thank Fabrizio Grosa, Felix Schlepper, and Chunzheng Wang, introducing me to the O^2 framework, helping me out when things seemingly did not make sense and implementing new features properly.

Thanks to Jacob Bastian van Beelen for machining all the parts 'ideally needed right now', and generally the mechanics and electronics teams.

Thank you to the OPMD team at Oxford, around Richard Plackett and Daniel Hynds, for guiding my first encounters with silicon sensors. To my fellow DPhils, especially Simon Koch, Maggie Chen, and Alessandro Ruggiero – thank you for the coffee breaks, lighthearted chats, and shared DWB endeavours. Importantly, also to our admin team around Sue Geddes, Kim Proudfoot, and Julieta Estremadoyro-Vermejo for working magic in all bureaucratic aspects, and enduring the modus operandi of never submitting anything before the deadline.

My time spent at Oxford will leave a lasting impact, as a truly unique place with Oriel College and the friends made there shaping the experience – including rowing in the freezing dark, winning blades on three occasions, and beating our Cambridge sister-college in Porto. To everyone who made my time at CERN and Oxford a unique experience – thank you!

Claire, you showed me what passion feels like and are the reason for many of my fondest memories. You ensured that I realise there is more to life beyond the DPhil – the experiences we shared will stay close to my heart. I am grateful we got to go through some of the most intense periods together while joking in the most unfiltered ways. Thank you for reminding me to embrace the *struggle as nature's way of strengthening it*, never stop aiming high and daring to dream bigger, both professionally and in my personal life.

To David – thank you for being what I imagine a brother to be like, helping me through various periods of life, always having an ear, it feeling like yesterday when we haven't heard each other for a while, and keeping me grounded with the occasional well-placed *reiss dich zam*.

Finally, to my friends and my family, thank you for being there and supporting me whenever I do reach out, tolerating that it is on unpredictable time scales – knowing you are here gives me the freedom to focus on the journey, which I value deeply.