# Memristors for classical and quantum applications

## Materials, devices, machine learning

Hua, E.

**DOI**
10.4233/uuid:ecf475d2-148a-4264-b949-9b66ff86ab2c

**Publication date**
2025

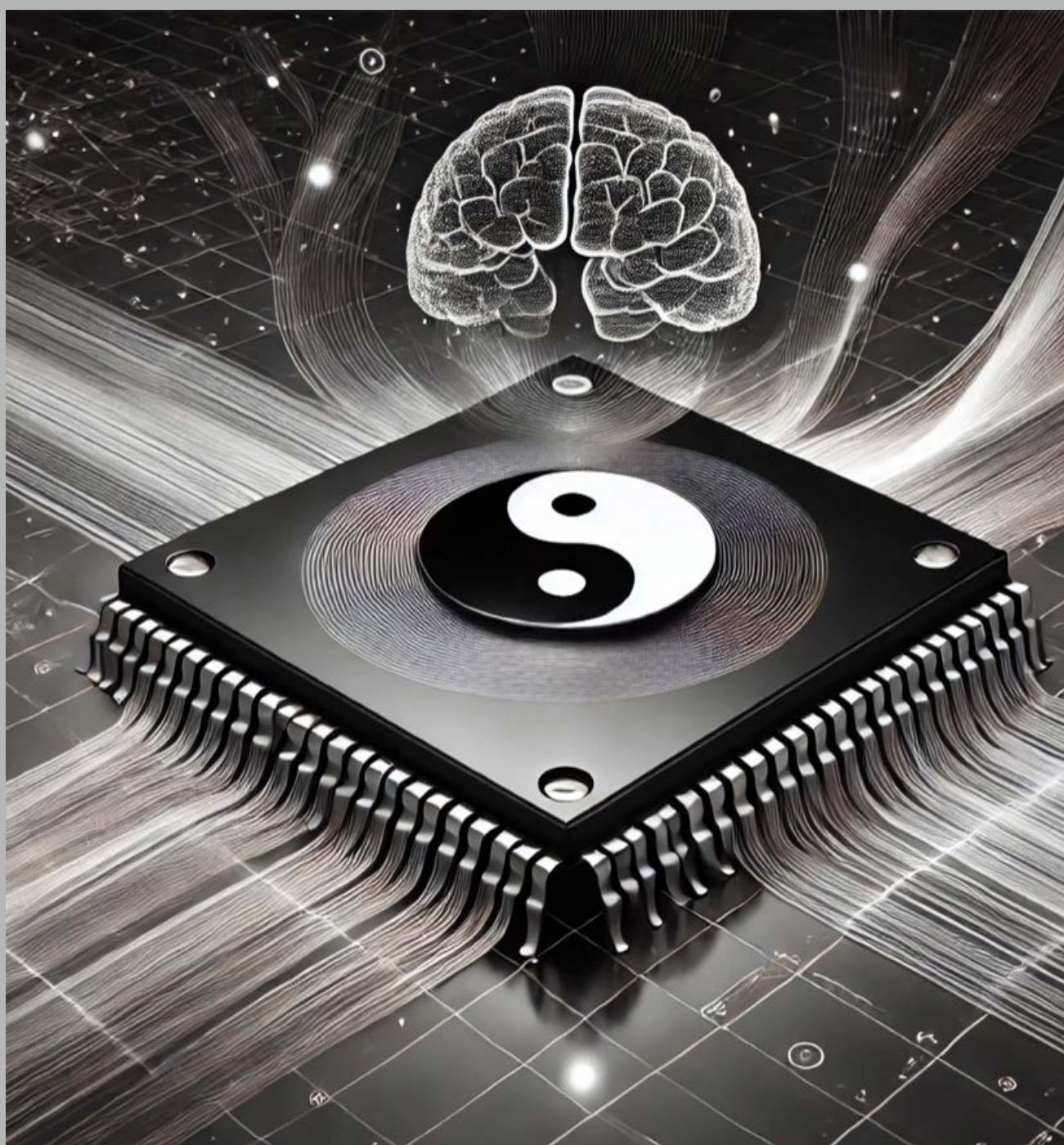
**Document Version**
Final published version

**Citation (APA)**
Hua, E. (2025). *Memristors for classical and quantum applications: Materials, devices, machine learning*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ecf475d2-148a-4264-b949-9b66ff86ab2c

# Memristors for Classical and Quantum Applications

MATERIALS, DEVICES, MACHINE LEARNING

**Erbing Hua**



# Propositions

Accompanying to the Ph.D. dissertation

**Memristors for Classical and Quantum Applications：**

**Materials, Devices, Machine learning**

By

**Erbing Hua**

1. Memristors are a promising candidate for brain-inspired synapses and neurons, though they alone cannot enable robotic consciousness. [This thesis]
2. Spiking neural networks encode information mainly through connectivity and spatial topology, rather than precise timing differences. [This thesis]
3. The future of quantum computing may hinge less on adding qubits and more on rethinking control—cryo-memristors show that progress requires both new devices and new mindsets. [This thesis]
4. Power laws emerge across domains, from ReRAM filament formation to citation networks, suggesting universal organizing patterns in both physics and society.
5. Technologies may eventually alleviate resource scarcity, but never the mental scarcity— the true origin of both cooperation and conflict in humanity.
6. Language shapes our understanding boundary across scientific disciplines; while each field develops its own vocabulary, all attempt to describe the same underlying reality from different perspectives.
7. Education should cultivate the ability to explore and generate new knowledge rather than simply transmit what is already known.
8. A true proponent must expose an idea's flaws more incisively than its critics to legitimately champion it.
9. Developing humanoid AGI robots is costly and risky; might it be wiser to invest in boosting human fertility instead?
10. Rituals encode cultural imperatives and promote conformity, just as a wedding signals the expectation of lifelong loyalty.

**These propositions are regarded as opposable and defendable, and have been approved as such by the promotors Dr. R. Ishihara and Prof. dr. ir. S. Hamdioui.**

# MEMRISTORS FOR CLASSICAL AND QUANTUM APPLICATIONS

MATERIALS, DEVICES, MACHINE LEARNING

# MEMRISTORS FOR CLASSICAL AND QUANTUM APPLICATIONS

MATERIALS, DEVICES, MACHINE LEARNING

## Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 3 November 2025 at 12:30 o'clock

by


## Erbing HUA

Master of Science in Chemistry, Shanghai Tongji University, China
born in Zhoukou, Henan, China

# CONTENTS

# SUMMARY

FROM the first spark of inspiration to the final forward-looking horizon, this thesis unfolds as a journey to re-imagine the foundations of computation. We merge breakthroughs in materials science, electronic device engineering, and deep generative learning to confront three of modern computing grandest challenges: the energy inefficiencies of classical architectures, the scaling limitations of neuromorphic hardware, and the exponential complexity of quantum systems.

We begin by identifying a threefold bottleneck at the heart of contemporary information processing. On one hand, the von Neumann architecture separates memory from logic, incurring high energy and latency costs. On another, quantum systems, with their exponentially expanding state spaces, defy conventional methods of characterization and control. Bridging these extremes demands both new materials and new paradigms: architectures that think and learn within memory itself and operate seamlessly across room-temperature and cryogenic domains. Our mission is to forge a unified computing framework that fuses neuromorphic principles, cryogenic and room-temperature memristors, and machine intelligence with quantum state tomography (QST).

The conceptual groundwork follows. Inspired by biological neurons, we explore how computation and memory can coexist within memristive architectures. Memristors, particularly resistive switching devices such as $HfO_2$-based ReRAM, emulate synaptic plasticity, enabling analog tuning and in-memory processing. We investigate both spiking and non-spiking neural models, contextualizing their use in QST. The core idea of computation-in-memory (CiM) emerges, performing neural operations directly within dense memristor crossbars, bypassing the von Neumann bottleneck. This unifying concept becomes the architectural backbone of our hybrid classical–quantum platform.

Our theoretical framework spans silicon physics, memristive mechanisms, and the formalism of quantum state reconstruction. We dissect electron-beam-induced processing (EBIP) as a route to room-temperature silicon device fabrication. We examine the physics of OxReRAM switching, ion migration, interfacial engineering, and energy barriers, and we extend this understanding to cryogenic regimes. In parallel, we articulate the formal structure of QST, density matrices, POVMs, and data scaling as $4^n$ for $n$ qubits, where neural networks emerge as natural generative or inference engines. Variational autoencoders, especially spiking VAEs (SVAEs), form the probabilistic bridge between neuromorphic learning and quantum reconstruction. The materials narrative begins with innovation in silicon processing. Abandoning high-temperature furnaces, we deploy spin-coated liquid polysilanes and transform them into functional amorphous silicon films via focused EBIP. *STEM–EELS* imaging, residual-gas analysis, and electrical characterization confirm uniform, low-defect films exhibiting stable ohmic behavior over months. This approach enables nanoscale precision and compatibility with flexible substrates, key for next-generation neuromorphic hardware.

ReRAM devices, long hampered by high-voltage electroforming and poor uniformity, are re-engineered. By designing Pd/$HfO_2$ interfaces, we realize forming-free OxReRAM cells that switch at sub-2 V, support multibit states, and retain data over $10^4$ s. Atomic-scale analysis reveals a Pd–O–Hf interfacial layer that stabilizes low-bias conductive pathways. These devices achieve endurance and energy consumption in the picojoule range, validating them as efficient synaptic elements for in-memory computing. At cryogenic temperatures, the same memristive principles enable a new frontier: **Cryo-Memristors for spin-qubit control**. Operating reliably at 4 K, Pt/Ti/$HfO_2$-based memristors and their modified variants (M-PtHT) serve as low-noise, multi-bit programmable gain elements for scalable quantum control electronics. Embedded near the quantum layer, these devices synthesize analog bias voltages with sub-100 µV resolution, drastically reducing the wiring complexity, heat load, and latency in large-scale qubit arrays. Statistical analysis shows linear resistance variation and stable multi-bit retention even at 4 K, confirming their potential as cryogenic analog memory elements for autonomous qubit tuning and adaptive quantum feedback. This chapter bridges device physics and quantum hardware, demonstrating that memristive programmability can extend beyond neuromorphic computing into the quantum domain.

We confront QST through the lens of machine learning. A diverse suite of neural architectures, FCN, CNN, RNN, RBM, CGAN, and Transformer, is deployed to reconstruct quantum states from simulated measurement data. Among them, CNNs deliver the best trade-off between fidelity and computational time, especially under expectation-based measurements. Yet, the SVAE architecture marks a turning point: as a generative probabilistic model, it achieves high-fidelity reconstructions even under sparse and noisy data, generalizing to higher qubit counts (up to 8) and scaling sub-exponentially in runtime. Its latent-space encoding of high-dimensional quantum information renders it ideal for real-time, energy-efficient inference when implemented on memristive crossbars. Simulations incorporating real device characteristics confirm that our forming-free and cryogenic OxReRAM-based CiM arrays can physically sustain deep QST networks. Memristor crossbars perform rapid in-memory matrix–vector multiplications, reducing inference energy by orders of magnitude compared to digital processors. Together, these results establish a scalable, hardware-aware path toward **hybrid classical–neuromorphic–quantum computing**.

We conclude by reflecting on the broader implications. This work demonstrates that room-temperature EBIP enables sustainable silicon fabrication; that forming-free OxReRAM devices can be engineered for reliable analog switching; that Cryo-Memristors enable scalable, low-power qubit control; and that generative neural networks, especially SVAEs, offer a pathway to efficient, hardware-embedded quantum state reconstruction. Looking ahead, these innovations converge toward cryogenic integration with quantum processors, adaptive quantum feedback via spiking neuromorphic circuits, and the eventual realization of intelligent, energy-aware quantum systems.

Through every chapter, one theme resounds: the dissolution of boundaries, between memory and logic, between classical and quantum, between matter and model. This thesis lays the foundation for a new kind of computing, one that learns like the brain, reasons like a physicist, and computes like the future demands.

# SAMENVATTING

VAN de eerste vonk van inspiratie tot aan de toekomstgerichte horizon ontvouwt dit proefschrift zich als een reis om de fundamenten van de berekening opnieuw te verbeelden. We combineren doorbraken in materiaalkunde, elektronische apparaatengineering en diep-generatief leren om drie van de grootste uitdagingen van de moderne informatietechnologie aan te pakken: de energie-inefficiëntie van klassieke architecturen, de schaalbaarheidsbeperkingen van neuromorfe hardware en de exponentiële complexiteit van kwantumsystemen.

We beginnen met het identificeren van een drievoudige flessenhals in hedendaagse informatieverwerking. Enerzijds scheidt de von Neumann-architectuur geheugen van logica, wat leidt tot hoge energie- en latentieverliezen. Anderzijds onttrekken kwantumsystemen, met hun exponentieel uitbreidende toestandsruimten, zich aan conventionele methoden van karakterisering en controle. Het overbruggen van deze uitersten vereist zowel nieuwe materialen als nieuwe paradigma's: architecturen die binnen het geheugen zelf kunnen denken en leren, en die naadloos functioneren over kamertemperatuur- en cryogene domeinen. Onze missie is het ontwikkelen van een verenigd rekenkader dat neuromorfe principes, cryogene en kamertemperatuurmemristoren, en machinale intelligentie samenbrengt met kwantumtoestandstomografie (QST).

De conceptuele basis volgt hieruit. Geïnspireerd door biologische neuronen onderzoeken we hoe berekening en geheugen kunnen samengaan binnen memristieve architecturen. Memristoren, met name resistieve schakelaars zoals $HfO_2$-gebaseerde ReRAM, imiteren synaptische plasticiteit en maken analoge afstemming en berekening in het geheugen mogelijk. We bestuderen zowel spiking- als niet-spikingneurale modellen en plaatsen hun gebruik in de context van QST. Het kernidee van computation-in-memory (CiM) komt hierbij naar voren: het uitvoeren van neurale berekeningen direct binnen dichte memristorkruisverbanden, waarmee de von Neumann-flessenhals wordt omzeild. Dit verenigende concept vormt de architectonische ruggengraat van ons hybride klassiek, kwantumplatform.

Ons theoretisch kader bestrijkt de fysica van silicium, memristieve mechanismen en de formele structuur van kwantumtoestandreconstructie. We analyseren elektronbundelgeïnduceerde processen (EBIP) als een route naar siliciumfabricage bij kamertemperatuur. We onderzoeken de fysica van OxReRAM-schakelingen, ionenmigratie, interfaciale engineering en energiebarrières, en breiden dit begrip uit naar cryogene omstandigheden. Parallel hieraan formuleren we de formele structuur van QST, dichtheidsmatrices, POVM's en de gegevensschaal als $4^n$ voor $n$ qubits, waarbij neurale netwerken optreden als natuurlijke generatieve of inferentie-instrumenten. Variational autoencoders, in het bijzonder spiking VAEs (SVAEs), vormen de probabilistische brug tussen neuromorf leren en kwantumreconstructie.

Het materiaalverhaal begint met innovatie in siliciumverwerking. Door af te zien van

hoge-temperatuurovens gebruiken we spin-gecoate vloeibare polysilanen en zetten deze om in functionele amorfe siliciumfilms via gerichte EBIP. STEM–EELS-beelden, restgasanalyse en elektrische karakterisering bevestigen uniforme, defectarme films met stabiel ohms gedrag gedurende maanden. Deze aanpak biedt nanometerschaalprecisie en compatibiliteit met flexibele substraten, essentieel voor neuromorfe hardware van de volgende generatie.

ReRAM-apparaten, die lang werden gehinderd door hoge vormingsspanningen en beperkte uniformiteit, worden herontworpen. Door Pd/$HfO_2$-interfaces te ontwerpen realiseren we vormingsvrije OxReRAM-cellen die schakelen bij minder dan 2 V, meerdere bittoestanden ondersteunen en gegevens behouden gedurende $10^4$ s. Atomair schaalonderzoek toont een stabiele Pd–O–Hf-interfacelaag die lage-biasgeleidende paden stabiliseert. Deze apparaten vertonen uitstekende duurzaamheid en energieverbruik in het picojoulebereik, waardoor ze worden gevalideerd als efficiënte synaptische elementen voor berekening in het geheugen.

Bij cryogene temperaturen maken dezelfde memristieve principes een nieuw grensgebied mogelijk: **Cryo-Memristoren voor spin-qubitbesturing**. Betrouwbaar werkend bij 4 K fungeren Pt/Ti/$HfO_2$-gebaseerde memristoren en hun gemodificeerde varianten (M-PtHT) als laagruisende, multibit programmeerbare versterkingselementen voor schaalbare kwantumregel-elektronica. Geïntegreerd nabij de kwantumlaag genereren deze apparaten analoge biasspanningen met een resolutie onder 100 µV, wat de bedradingcomplexiteit, warmtebelasting en latentie in grootschalige qubitarrays drastisch vermindert. Statistische analyse toont lineaire weerstandvariatie en stabiele multibitretentie, zelfs bij 4 K, wat hun potentieel bevestigt als cryogene analoge geheugenelementen voor autonome qubitafstemming en adaptieve kwantumterugkoppeling. Dit hoofdstuk overbrugt de fysica van apparaten en kwantumhardware en toont aan dat memristieve programmeerbaarheid verder reikt dan neuromorfe berekening, tot in het kwantumdomein.

We benaderen QST door de lens van machinaal leren. Een breed scala aan neurale architecturen, FCN, CNN, RNN, RBM, CGAN en Transformer, wordt ingezet om kwantumtoestanden te reconstrueren uit gesimuleerde meetgegevens. Onder deze leveren CNN's de beste balans tussen nauwkeurigheid en rekentijd, vooral bij verwachting-gebaseerde metingen. Toch markeert de SVAE-architectuur een keerpunt: als generatief probabilistisch model behaalt zij hoge-fideliteitsreconstructies, zelfs onder schaarse en ruisachtige data, generaliserend tot hogere qubitaantallen (tot 8) en schalend sub-exponentieel in rekentijd. De latente-ruimtecoderingscapaciteit van hoge-dimensionale kwantuminformatie maakt dit model ideaal voor realtime, energie-efficiënte inferentie wanneer geïmplementeerd op memristorkruisverbanden. Simulaties waarin reële apparaatkarakteristieken zijn opgenomen, bevestigen dat onze vormingsvrije en cryogene OxReRAM-gebaseerde CiM-arrays diepe QST-netwerken fysiek kunnen ondersteunen. Memristorkruisverbanden voeren snelle matrix–vectorvermenigvuldigingen in het geheugen uit, wat het energieverbruik met orders van grootte verlaagt ten opzichte van digitale processoren. Gezamenlijk vestigen deze resultaten een schaalbaar, hardware-bewust pad naar **hybride klassiek–neuromorf–kwantumrekenen**.

Tot slot reflecteren we op de bredere implicaties. Dit werk toont aan dat EBIP bij kamertemperatuur duurzame siliciumfabricage mogelijk maakt; dat vormingsvrije OxReRAM-apparaten betrouwbaar analoog schakelen; dat Cryo-Memristoren schaalbare, energie-

zuinige qubitbesturing mogelijk maken; en dat generatieve neurale netwerken, met name SVAEs, een pad bieden naar efficiënte, in hardware ingebedde kwantumtoestandreconstructie. Vooruitkijkend convergeren deze innovaties naar cryogene integratie met kwantumprocessors, adaptieve kwantumterugkoppeling via spiking-neuromorfe circuits en de uiteindelijke realisatie van intelligente, energie-bewuste kwantumsystemen.

Door elk hoofdstuk heen klinkt één thema door: het oplossen van grenzen, tussen geheugen en logica, tussen klassiek en kwantum, tussen materie en model. Dit proefschrift legt het fundament voor een nieuw soort berekening, één die leert als het brein, redeneert als een natuurkundige en rekent zoals de toekomst vereist.

# List of Figures

# List of Tables

# 1

# INTRODUCTION

1

## 1.1. RESEARCH BACKGROUND

THE evolution of computing has been marked by transformative breakthroughs that have reshaped society and science alike [1]. From the early days of vacuum tube computers to today's advanced multi-core processors, the quest for higher performance, efficiency, and scalability has driven innovation across multiple disciplines [2–5]. This section provides a broad overview of the key trends, challenges, and emerging paradigms that underpin modern computing research.

Traditional digital computing systems, primarily based on the *von Neumann* architecture, revolutionized information processing by introducing a clear separation between memory and processing units [2, 6]. This architecture enabled tremendous advances in computer science and engineering, facilitating the development of complex algorithms and large-scale data processing. However [3, 7, 8], as data volumes have exploded due to advances in sensor technologies, the Internet of Things (IoT), and big data analytics, the inherent limitations of the *von Neumann* paradigm have become increasingly apparent. The so-called *von Neumann* bottleneck, the energy and time costs associated with shuttling data between separate memory and processing units, now poses a significant constraint on further performance improvements.

In response to these limitations, researchers have explored alternative computing paradigms that mimic the highly parallel and energy-efficient processing capabilities of biological systems. Neuromorphic computing, for instance, draws inspiration from the structure and function of the human brain [9]. By emulating neural architectures and synaptic dynamics, neuromorphic systems offer the potential for real-time, low-power processing [10]. Early work by pioneers such as Carver Mead laid the groundwork for neuromorphic engineering, introducing the idea that analog and mixed-signal circuits could be designed to replicate the brain's computational efficiency [9]. Today, neuromorphic research spans a wide array of topics, from spike-based processing in spiking neural networks (SNNs) to the development of novel hardware components that mimic synaptic plasticity [10–13].

A critical component of neuromorphic computing is the development of emerging memory devices that can simultaneously store and process information. Traditional *CMOS*-based memories, while highly reliable, are often not optimized for the energy efficiency and parallelism required by next-generation applications. In contrast, devices such as memristors [14–16], including resistive random-access memories (ReRAMs), exhibit analog tunability and non-volatile behavior that make them promising candidates for in-memory computing [17]. Their ability to modulate resistance in response to electrical stimuli offers a hardware means to implement learning rules akin to biological synapses, potentially enabling dense, energy-efficient neural networks.

At the same time, the landscape of computing research is being transformed by the rapid development of quantum technologies. Quantum computing promises to tackle problems intractable for classical systems by exploiting phenomena such as superposition and entanglement. However, the very nature of quantum systems introduces a new set of challenges. For example, quantum state characterization through quantum state tomography (QST) is fundamentally limited

by the exponential growth of the state space with the number of qubits [18–20]. Conventional QST methods, while effective for small systems, quickly become impractical as system size increases, prompting researchers to explore alternative strategies that combine insights from machine learning and advanced hardware architectures.

One particularly pressing challenge lies in the integration of quantum hardware with classical control systems. As the number of qubits grows, traditional wiring architectures, based on individual digital-to-analog converters and signal lines, become unsustainable due to increased thermal loads and wiring complexity. This scaling bottleneck poses a serious obstacle to the deployment of large-scale quantum processors.

Recent research has highlighted the potential of cryo-compatible memristors as an innovative solution to this challenge. Operating reliably at temperatures as low as 4 K, cryogenic memristors can function as programmable gain elements or memory-inference engines near the quantum layer [21]. By synthesizing analog voltages directly at cryogenic stages, these devices reduce the number of control lines required, minimize heat dissipation, and enable scalable, energy-efficient quantum-classical interfacing. Their multi-bit programmability, retention stability, and low power operation under cryogenic conditions offer promising opportunities for in-situ learning and adaptive quantum feedback control.

The growing convergence between classical, neuromorphic, and quantum computing has thus spurred interest in hybrid approaches that leverage the strengths of each domain [22, 23]. In these hybrid systems, novel materials and low-energy devices are integrated with machine learning algorithms to overcome the limitations imposed by traditional architectures [24]. For example, integrating computation-in-memory (CiM) techniques into neuromorphic systems can dramatically reduce the energy overhead associated with data transfer [25]. By performing matrix–vector multiplications directly within memory arrays, these architectures address one of the most significant bottlenecks in conventional systems and open new avenues for efficient, real-time processing.

Moreover, advances in nanofabrication and material science continue to play a pivotal role in pushing the boundaries of what is technologically possible [26, 27]. Low-temperature processing methods, enabled by innovations in chemical synthesis and electron-beam processing, are emerging as attractive alternatives to traditional high-temperature fabrication techniques [28]. These methods not only reduce energy consumption but also expand the range of compatible substrates, paving the way for flexible and wearable electronics.

The broad research landscape in modern computing is characterized by an urgent need to overcome the limitations of existing architectures. The challenges include mitigating the energy inefficiencies of data transfer in conventional systems, harnessing the parallelism and adaptability of neuromorphic computing, scaling quantum technologies to practical levels, and embedding efficient control mechanisms at cryogenic temperatures. Researchers are now exploring interdisciplinary solutions that combine breakthroughs in materials science, advanced nanofabrication, emerging memory devices, cryogenic electronics, and machine

learning. This integrated approach holds the promise of developing a new generation of computing systems that are not only more powerful and efficient but also capable of addressing the complex, data-intensive challenges of the 21st century.

## 1.2. MOTIVATION

THE relentless miniaturization and energy constraints in modern computing have spurred the search for unconventional architectures that break away from the traditional *von Neumann paradigm*. In particular, neuromorphic computing, an approach that emulates the architecture and dynamics of the human brain, has attracted considerable attention as it promises massive parallelism, low power consumption, and real-time adaptive processing. Concurrently, the challenges inherent in quantum state tomography (QST) for characterizing increasingly complex quantum systems have motivated the development of efficient, hardware-based computational schemes. The work presented in this thesis integrates these two domains by leveraging emerging memristive technologies to construct neuromorphic hardware that can perform QST with unprecedented energy efficiency and scalability.

This part details motivation for the research, drawing from the experimental and theoretical advancements described in Chapters 4, 5, 6, and 7. In Chapter 4, novel materials and device fabrication methods based on electron-beam-induced processing (EBIP) and liquid-phase silicon (L-Si) precursors are introduced as a means to overcome the high-temperature and contamination issues plaguing traditional silicon processing. Chapter 5 extends this work into the realm of resistive switching memories by exploring forming-free OxReRAM devices that utilize $HfO_2$-based dielectrics and palladium (Pd) electrodes to achieve low-voltage, multibit switching. Chapter 6 shifts the focus toward quantum-compatible electronics, presenting cryogenic memristors capable of operating reliably at 4 K. These devices are envisioned as programmable analog elements for *in-situ* quantum control, addressing the wiring and thermal bottlenecks that hinder scalability in quantum-dot and spin-qubit platforms. By enabling compact, low-power voltage biasing circuits near the quantum layer, cryo-memristors open up new directions for efficient integration of classical neuromorphic logic with quantum processors. Finally, Chapter 7 explores the application of these advanced devices as the hardware substrate for neuromorphic neural networks that perform QST, thereby addressing the "curse of dimensionality" that hinders classical QST methods.

The motivation for this research is fourfold. First, there is a pressing need to develop new materials and device architectures that can deliver the energy efficiency and high-density integration required by next-generation computing systems. Second, the development of forming-free resistive switching devices is essential to overcome the limitations of conventional ReRAM technologies, such as high electroforming voltages and significant device-to-device variability. Third, by integrating these novel devices into neuromorphic networks, it becomes possible to tackle large-scale quantum state reconstruction, a critical challenge in the era of quantum computing. Fourth, the demonstration of cryo-compatible memristors presents a promising solution to the wiring complexity and energy dissipation

challenges faced in scaling quantum control electronics. Their ability to support low-temperature, multi-level analog programmability positions them as key enablers for future hybrid quantum–classical systems with embedded intelligence.

### 1.2.1. ADVANCED MATERIALS AND DEVICE FABRICATION

Traditional silicon processing techniques, such as chemical vapor deposition (CVD) and molecular beam epitaxy (MBE), typically require high temperatures and complex setups that limit substrate compatibility and introduce unwanted impurities [29]. The high processing temperatures are not only energy intensive but also incompatible with a range of flexible and temperature, sensitive substrates. Furthermore, the presence of contaminants during these high-temperature processes can detrimentally affect the performance and reliability of semiconductor devices. To overcome these challenges, the present work introduces an innovative approach that employs liquid-phase silicon (L-Si) precursors, specifically polysilanes, as an alternative to traditional solid-state precursors. Polysilanes can be easily spin-coated onto various substrates and subsequently converted into silicon films through low-temperature electron-beam-induced processing (EBIP) [30, 31]. EBIP offers nanoscale resolution while operating at room temperature, thereby eliminating the need for high-temperature annealing. This technique not only reduces energy consumption but also minimizes contamination, as the process can be carried out under a controlled inert atmosphere. The successful demonstration of EBIP for fabricating high-quality silicon nanostructures opens up new avenues for the scalable integration of advanced semiconductor devices [31]. The ability to process silicon at room temperature using EBIP enables the fabrication of devices on a much broader range of substrates, including flexible and polymer-based materials. This versatility is critical for the development of next-generation electronics, where integration density and low power consumption are of paramount importance. Moreover, the high resolution offered by EBIP allows for precise control over device dimensions, which is essential for minimizing leakage currents and optimizing the performance of nanoscale devices.

### 1.2.2. MEMRISTOR DEVICES

ReRAM devices have been widely studied for their potential in non-volatile memory and neuromorphic computing applications [32, 33]. However, conventional ReRAM devices typically require an electroforming step, a high-voltage pre-process that initiates the formation of conductive filaments (CFs) in the resistive medium. This electroforming process not only consumes significant energy but also introduces variability and reliability issues, which limit device endurance and hinder large-scale integration. Recent research has demonstrated that by carefully engineering the dielectric layer, it is possible to suppress the need for electroforming altogether. In particular, hafnium oxide ($HfO_2$) has emerged as a promising dielectric material due to its high breakdown voltage, thermal stability, and compatibility with complementary metal-oxide-semiconductor (CMOS) processes [32–35]. Furthermore, the incorporation of Multi-Level Resistance (MLR) technology in $HfO_2$ boosts

memory density and slashes manufacturing costs, offering significant advantages for neuromorphic computing applications. While the problem is there that the high voltage of the electroforming process is energy-intensive. In order to resolve the high voltage of the electroforming process, multiple techniques have been applied to increase the defective states in the dielectric materials so that the voltages of generating the conductive pathways are lower or eliminated, including X-ray irradiation, high-temperature thermal annealing, exotic element doping, etc. [36–40].Therefore, there is a compelling need for simpler, energy-efficient fabrication processes for forming-free ReRAM devices. From the above techniques, the dielectric material $HfO_2$ shows high capacity for modification as well. The integration of forming-free OxReRAM devices into neuromorphic circuits offers a transformative route toward energy-efficient computing. The elimination of the electroforming step reduces energy overhead, while the inherent multibit capability of these devices enables the fine-tuning of synaptic weights in artificial neural networks. By implementing these devices in crossbar arrays, it is possible to perform matrix–vector multiplications (MVMs) directly in memory, thereby bypassing the energy-hungry data transfer bottleneck inherent in conventional architectures [41].

### 1.2.3. CRYOGENIC MEMRISTORS FOR QUANTUM CONTROL

The scalability of quantum processors based on spin qubits or quantum dots is critically limited by the complexity of cryogenic control wiring and the associated thermal load. Conventional control systems rely on room-temperature digital-to-analog converters (DACs) and multiplexed wiring to bias each qubit individually, resulting in significant parasitic heat transfer and a lack of local adaptability [42, 43]. As quantum devices scale to hundreds or thousands of qubits, these issues become increasingly detrimental, necessitating innovations in cryo-compatible electronics for efficient control, configuration, and feedback.

To address this challenge, recent advances have demonstrated the feasibility of using memristive devices as analog control elements operating at cryogenic temperatures [44]. In particular, resistive switching devices based on $HfO_2$ and other transition metal oxides have exhibited stable multibit programming behavior at 4 K, making them suitable candidates for integration into low-temperature control circuits. These cryogenic memristors retain their analog tunability and non-volatile memory function even under ultra-low temperature operation, allowing them to serve as programmable gain elements or voltage bias sources directly adjacent to quantum devices.

One promising approach involves the use of memristive cryo-DACs to locally generate analog voltages for qubit gate control, replacing bulky room-temperature electronics and significantly reducing the number of required signal lines [44, 45]. This strategy not only minimizes wiring complexity but also reduces energy dissipation, as programming currents are confined to sub-nanoampere levels and updates can be made infrequently. Furthermore, by embedding memristive weights directly into the cryogenic environment, it becomes possible to envision closed-loop learning systems for adaptive qubit calibration, machine learning-based error correction, and scalable quantum–classical interfacing [46].

The compatibility of cryogenic memristors with standard CMOS back-end-of-line (BEOL) integration further supports their incorporation into 3D-stacked quantum computing architectures [47]. Their small footprint, low-power requirements, and analog programmability make them ideal building blocks for future hybrid systems that combine classical inference engines and quantum logic in close physical proximity. These advances suggest a compelling path forward for building highly scalable, energy-efficient quantum processors with embedded control logic.

### 1.2.4. MEMRISTOR-COMPATIBLE NEURAL NETWORKS FOR QST

QST, A promising approach to reconstruct quantum system: QST is the process of reconstructing the full quantum state of a system based on measurement data. As quantum systems scale up in complexity, traditional QST methods face an exponential increase in the number of parameters that must be estimated. For an $N$-qubit system, the density matrix is a $2^N \times 2^N$ object that requires the estimation of $4^N - 1$ independent parameters [48]. This "curse of dimensionality" makes conventional QST computationally prohibitive for systems beyond a few qubits.

Machine Learning Approaches to Quantum State Reconstruction: Recent advances in machine learning have offered promising new avenues for overcoming the scalability issues associated with QST. Neural network architectures, including fully connected networks (FCNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), variational autoencoders (VAEs), and generative adversarial networks (GANs), have been successfully applied to the problem of reconstructing quantum states from incomplete and noisy measurement data [49]. These methods learn to map high-dimensional measurement data directly to the corresponding density matrix, effectively bypassing the need for iterative, resource-intensive optimization routines.

Integration of Memristor-Compatible Computation-in-Memory (CiM): A critical innovation is the integration of memristor-based computation-in-memory (CiM) architectures with neural network models for QST. Memristors, which inherently combine memory and processing functionalities, are ideally suited for executing the matrix–vector multiplications that underpin neural network operations. By performing these computations directly in memory, CiM architectures eliminate the energy and latency penalties associated with data transfer in traditional *von Neumann* systems. This integration is particularly advantageous for QST, where large-scale, high-dimensional data must be processed rapidly and energy efficiently [50].

### 1.2.5. SYNERGY AND INTERDISCIPLINARY IMPACT

Bridging the Gap Between Classical and Quantum Domains: The work presented herein represents a convergence of three rapidly evolving research fields: advanced materials and device fabrication, non-volatile memory technologies for neuromorphic computing, and QST. By addressing the critical challenges in each domain, from the high-temperature constraints of silicon processing to the scalability issues in QST, the proposed research establishes a framework that bridges the gap between classical electronic systems and emerging quantum technologies.

Energy Efficiency and Sustainable Computing: One of the most compelling motivations behind this research is the drive toward sustainable, energy-efficient computing. As data centers and high-performance computing systems continue to consume vast amounts of energy, there is an urgent need for architectures that can deliver high performance without incurring excessive power costs. Neuromorphic systems inspired by the brain's unparalleled efficiency offer a viable alternative, and the integration of forming-free ReRAM devices further enhances this promise by reducing the energy overhead associated with memory operations.

Scalability and Real-Time Processing: The ability to scale computation to accommodate the demands of large-scale quantum systems is another critical driver of this work. Traditional approaches to QST become intractable as the number of qubits increases, but by leveraging memristor-based neural networks and CiM architectures, it becomes possible to perform quantum state reconstruction in real time. This capability is not only essential for the practical deployment of quantum processors but also paves the way for real-time quantum error correction and adaptive quantum control.

Interdisciplinary Collaboration and Future Prospects: The interdisciplinary nature of this research, spanning materials science, electrical engineering, computer science, and quantum physics, highlights the importance of collaborative efforts in solving complex technological challenges. The novel device architectures and computational paradigms introduced in this thesis are expected to have far-reaching implications, influencing the development of next-generation quantum computing systems, low-power neuromorphic processors, and advanced sensor networks. Furthermore, the experimental methodologies and simulation frameworks developed herein provide a robust foundation for future research into hybrid quantum-classical architectures.

## 1.3. PROBLEM STATEMENTS

THIS thesis targets the device–to–system bottlenecks that limit reliable, energy-efficient memristive compute-in-memory (CiM) for quantum-adjacent workloads, with quantum state tomography (QST) as the running benchmark. Trade-offs and surveys appear once in Chapter 3; here we state the concrete problems, the guiding questions, and where each is addressed in the thesis. The identified problems are grouped into five categories, each corresponding to a subsequent experimental or computational chapter."

**Problem 1: Interface & forming control in $HfO_2$ ReRAM.** Stochastic electroforming and poorly shaped oxygen-vacancy profiles undermine multilevel linearity, repeatability, and energy. How can we eliminate electro-forming process while maintaining tunable, stable multilevel conductance suitable for analog MACs? Chapter 5 introduces interface-engineered Pd/$HfO_2$ stacks that are forming-free and support controlled conductance updates; we quantify D2D/C2C spread, programming energy, and linearity versus compliance/verify strategy.

**Problem 2: Cryogenic stability, transport regimes, and ultra-low noise.** From 300 K to cryogenic temperature, retention drift, $1/f$ noise, and read-disturb jeopardize fine multilevel control (needed for sub-mV, down to sub-100 $\mu$V, biasing in quantum control). Transport mechanisms can change with compliance and temperature, altering variability. *Guiding questions:* (i) Which bias, compliance, and dwell conditions minimize drift and noise at 4–77 K; (ii) Which transport regimes (Ohmic vs. trap-assisted) dominate under different compliance ranges and how do they impact level stability? Chapter 6 maps cryogenic operating windows, extracts drift exponents and noise PSD, and relates regime changes to programming conditions; we demonstrate stable, fine-step programming at cryo under low power.

**Problem 3: Model–hardware mismatch for QST inference.** Neural models trained under ideal arithmetic degrade on analog/non-ideal hardware (finite precision, IR-drop, read noise, drift). How should training and quantization be co-designed to absorb hardware errors without large accuracy loss? Chapter 7 introduces noise/quantization-aware training aligned to device granularity and peripheral limits; we compare CNN/RBM/CGAN/Transformer/SVAE under identical data and calibrated noise injectors.

**Problem 4: QST shot/sample complexity.** Static measurement policies scale poorly with qubit number and available shots. Can uncertainty-aware (active) acquisition reduce shots at fixed fidelity and remain robust to hardware noise? Chapter 7 evaluates active vs. static schemes and shows models (esp. SVAE) maintain fidelity under sparse/noisy shots when paired with active measurement selection.

**Problem 5: Lack of localized, BEOL-compatible actuation for maintenance.** There is no compact method to re-center arrays or reduce variability without global heating or disrupting cryo operation. Can we add a localized actuator that performs sub-$\mu$s conditioning/anneal within backend thermal budgets and cryo constraints? Chapter 4 introduces a BEOL-compatible liquid-silicon (*L–Si* module providing spatially confined thermal/electrical pulses that accelerate calibration and mitigate drift with negligible global thermal load.

**Research Questions:**

**RQ1. Forming-free operation (Ch. 5).** How can interface engineering ($Pd/HfO_2$) eliminate electroforming and yield stable, tunable multilevels with tight D2D/C2C dispersion and low programming energy?

**RQ2. Cryo physics & noise (Ch. 6).** What bias/compliance/dwell windows minimize drift and $1/f$ noise at 4–77 K, and how do transport regimes under different compliances affect fine-step stability and read-disturb?

**RQ3. Fine biasing at cryo (Ch. 6).** Which programming/verify protocols achieve sub-mV (target: sub-100 $\mu$V) stable steps over hours with acceptable duty cycle and power, enabling quantum-adjacent biasing?

**RQ4. Calibration for arrays (Ch. 5).** Which lightweight routines—adaptive write-verify, read-averaging, LUT linearization—maximize effective MAC accuracy under IR-drop, dispersion, and converter limits?

**RQ5. Robust QST inference (Ch. 7).** Which NN families (CNN vs. SVAE, etc.) and noise/quantization-aware training schemes preserve QST fidelity under analog precision limits, drift, and read noise?

**RQ6. Active acquisition (Ch. 7).** Can uncertainty-aware measurement selection reduce required shots at a fixed fidelity target and remain robust under device/circuit non-idealities?

**RQ7. Localized actuation (Ch. 4).** Can a BEOL-compatible *liq-Si* actuator provide spatially confined conditioning that shortens calibration, recenters drift, and respects backend/cryo thermal budgets?

**Reading map.** Background & metrics (Ch. 3); forming-free Pd/$HfO_2$ devices (Ch. 5); cryogenic stability/noise and fine-bias protocols (Ch. 6); array calibration and crossbar architecture (Ch. 5); models/training and active tomography (Ch. 7); *liq-Si* integration and HIL system validation (Ch. 4). These questions collectively frame the methodology and chapter organization described next.

## 1.4. STRUCTURE OF THE THESIS

This dissertation is structured into eight chapters that together form a coherent journey from materials science to device physics and system-level neuromorphic and quantum computing applications. The organization follows a bottom-up logic, from atomic-scale materials to system integration, reflecting how the individual elements of this research converge into a unified framework.

**Chapter 1** introduces the motivation, background, and research objectives. It defines the central problem of energy-inefficient and thermally constrained control in both classical and quantum systems and motivates the use of memristor-based computation-in-memory (CiM) to address these challenges. The chapter also summarizes the research questions and outlines the overall strategy connecting materials, devices, and neural models.

**Chapter 2** reviews the conceptual foundation of neuromorphic computing and its relevance to quantum information processing. It surveys memristor technologies as building blocks of CiM, discusses the operation principles of spiking and non-spiking neural networks, and analyzes machine-learning architectures, including FCN, CNN, RNN, RBM, CGAN, Transformer, and VAE, used for QST. The review establishes the theoretical bridge between memristive hardware and neural inference for quantum state reconstruction.

**Chapter 3** provides the theoretical background required for the interdisciplinary approach of this work. It explains the physical mechanisms of low-temperature silicon formation through electron-beam-induced processing (EBIP), the resistive switching mechanisms in $HfO_2$-based memristors, and the formalism of quantum state tomography based on POVMs and density matrices. Furthermore, it presents

the concept of probabilistic inference using spiking variational autoencoders (SVAEs), linking neural computation to quantum-state reconstruction.

**Chapter 4** focuses on materials and fabrication methods. Using spin-coated liquid silicon precursors and EBIP, amorphous silicon thin films are formed at room temperature without the need for high-temperature annealing. The process yields defect-poor and uniform silicon layers validated by RGA, HR-STEM, and EELS analyses. The chapter demonstrates that EBIP-based silicon is a viable low-temperature, BEOL-compatible route for future neuromorphic and flexible device platforms.

**Chapter 5** introduces the design and electrical characterization of forming-free $Pd/HfO_2$ memristors. Interface engineering using Pd results in the formation of a Pd–O–Hf layer that suppresses stochastic filament formation and enables stable multibit switching at sub-2 V. The chapter analyzes conduction mechanisms, retention, and endurance, supported by compact modeling and Cadence simulations. It demonstrates how these devices achieve picojoule-level switching energy and suitability for analog synaptic operation in spiking neural networks.

**Chapter 6** extends the device framework into the cryogenic regime. $Pt/Ti/HfO_2/Pt$ and modified M–PtHT devices are fabricated and characterized down to 4 K, showing reliable multibit operation and suppressed noise. Cryo-circuit simulations verify that these devices can serve as analog memory and programmable-gain elements for spin-qubit biasing, reducing interconnect complexity and power dissipation at milli-kelvin temperatures.

**Chapter 7** implements neural-network-based quantum state tomography (QST) on both conventional and memristor-compatible architectures. Various models, including FCN, CNN, RBM, CGAN, Transformer, and SVAE, are benchmarked with respect to fidelity, runtime, and energy consumption. Hardware-aware training, including noise injection and quantization, aligns the algorithmic inference with real device constraints. The chapter concludes by demonstrating that CGAN-, CNN- and SVAE-based QST offers the promising trade-off among energy, speed, and reconstruction accuracy.

**Chapter 8** concludes the thesis by synthesizing the results and highlighting future research directions. It discusses how the developed EBIP materials, forming-free memristors, and cryogenic analog memories jointly enable scalable, low-power control for hybrid classical–neuromorphic–quantum systems. The chapter closes with an outlook on integration challenges, reliability considerations, and the long-term vision of self-adaptive cryogenic control based on neuromorphic principles.

Overall, the thesis forms a continuous narrative, from materials discovery to algorithmic implementation, demonstrating how memristor technology can bridge classical, neuromorphic, and quantum computing paradigms.

## 1.5. CONTRIBUTIONS OF THIS THESIS

IN summary, the motivation for this thesis arises from the need to address the critical challenges at the intersection of energy efficiency, device scalability, and quantum information processing. The innovative approaches presented in

Chapters 4, 5, and 6, ranging from low-temperature EBIP fabrication of silicon nanostructures, through the development of forming-free OxReRAM devices, to the integration of memristor-based neural networks for QST, collectively pave the way toward a new generation of computing technologies. These advances promise not only to enhance the performance and reliability of neuromorphic systems but also to enable practical, real-time quantum state reconstruction for complex quantum systems.

The work presented here represents a significant step forward in overcoming the limitations of conventional semiconductor processing and memory technologies. By rethinking the fundamental architectures of electronic devices and integrating novel materials with cutting-edge computational models, this research lays the groundwork for energy-efficient, scalable systems that can meet the demands of future quantum and neuromorphic applications.

# 2

# NEUROMORPHIC DEVICES AND ANNS FOR QST

*QST, the task of reconstructing an unknown quantum state from measurement data, faces a notorious explosion of complexity as quantum system size grows. Conventional tomography methods demand exponentially many measurements and intensive classical post-processing, becoming intractable for high-dimensional states. Machine learning approaches have recently emerged as powerful alternatives: neural networks can serve as compact ansätze to represent quantum states and extract key properties from limited data In parallel, neuromorphic computing based on memristive devices offer a path to implement these neural networks in specialized hardware, with co-located memory and computation for massive parallelism and energy efficiency. This review provides a comprehensive survey of the intersection of these fields. We introduce the motivations for applying machine learning and neuromorphic hardware to QST, then review memristor device physics and their role as analog synapses in computing in memory . We outline neuromorphic architectures, contrasting non-spiking analog networks with spiking neural networks, and discuss their potential for efficient neural computing. We then survey a spectrum of neural network models used for QST, including Restricted Boltzmann Machines, CNNs, RNNs, GANs, Transformers, and VAEs, evaluating their performance and prospects for neuromorphic implementation. A comparative analysis of classical versus neuromorphic implementations is presented, focusing on scalability, energy efficiency, and accuracy. Finally, we discuss current limitations (device variability, training challenges, integration issues) and highlight open research directions toward integrating neuromorphic hardware with quantum information systems. Throughout, our focus remains on established results from peer-reviewed literature, aiming to provide a neutral, thorough assessment of this emerging multidisciplinary area.*

## 2.1. QST AND NEUROMORPHIC APPROACHES

QUANTUM state tomography is essential for validating and characterizing quantum systems, yet it becomes overwhelmingly difficult in high dimensions [24, 51]. The core challenge is the *curse of dimensionality*: an $N$-qubit pure state is described by a vector in a $2^N$-dimensional Hilbert space (or a density matrix with $(2^N)^2 - 1$ parameters for mixed states). Consequently, standard tomography requires a number of measurements and computational resources that grow exponentially with $N$[24, 51]. Indeed, brute-force methods like maximum likelihood estimation often become unfeasible except for very small systems [24]. As quantum simulators and processors enter the noisy intermediate-scale quantum (NISQ) era with tens or hundreds of qubits, new scalable tomography methods are needed [19, 51].

Machine learning (ML) for quantum state tomography (QST) has gained traction as a promising solution. Neural networks can serve as flexible function approximators to learn the mapping from measurement outcomes to an estimated quantum state. Unlike traditional approaches, ML-based tomography can exploit patterns in data to compress the representation of quantum states. For example, Torlai *et al.* demonstrated that neural networks can accurately reconstruct highly entangled states of over 100 qubits from relatively simple measurement data, far beyond the reach of brute-force methods [24]. Similarly, an adaptive tomography scheme using neural networks was shown to achieve orders-of-magnitude speed-ups in state reconstruction while retaining high accuracy [52]. The success of these approaches is underpinned by the ability of neural networks to approximate a wide variety of quantum states (leveraging universal approximation theorems) and to encode volume-law entanglement efficiently [51]. By training on measurement data, neural models can learn an implicit model of the quantum state, enabling tasks like predicting expectation values or entanglement measures directly [24].

However, software solutions alone may not suffice as quantum systems continue to grow. Even if a neural network compresses the state, training that network or inference (tomographic reconstruction) may be computationally intensive on classical hardware. This is where *neuromorphic hardware* enters the picture. Neuromorphic computing refers to non-von Neumann architectures inspired by the brain, which integrate memory and processing and operate in parallel across many simple *neurons* and *synapses* [53, 54]. In the context of QST, neuromorphic hardware could provide fast, scalable platforms to implement ML-based tomography algorithms in real-time or with minimal energy consumption [55, 56]. The idea is that a physical neural network (realized in electronics or other substrates) could process quantum measurement data and output an estimated state much faster than a conventional computer, enabling online tomography or adaptive measurement loops at the speed of the experiment. This is particularly relevant for quantum information systems where measurements might need to be processed on the fly (for example, for adaptive tomography or error diagnostics in quantum processors).

In this chapter, we examine the key ingredients and literature at this intersection. First, we review *memristor technology*, a class of resistive memory devices that underpin many neuromorphic architectures. We discuss memristors' physical principles, types, and how they enable computing in memory and synapse-like

behavior. Next, we survey neuromorphic architectures, distinguishing between non-spiking analog designs and spiking neural networks, highlighting their respective advantages for efficient computation. We then provide a comprehensive overview of neural network models applied to QST, from established approaches like Restricted Boltzmann Machines to cutting-edge transformers, and assess their compatibility with neuromorphic implementation. A comparative discussion of classical versus neuromorphic implementations of neural networks is given, focusing on scalability, energy efficiency, and accuracy as reported in the literature. Finally, we conclude with a critical discussion of current limitations and future prospects for integrating neuromorphic hardware into quantum information science, outlining research directions that could bridge the gap between these domains.

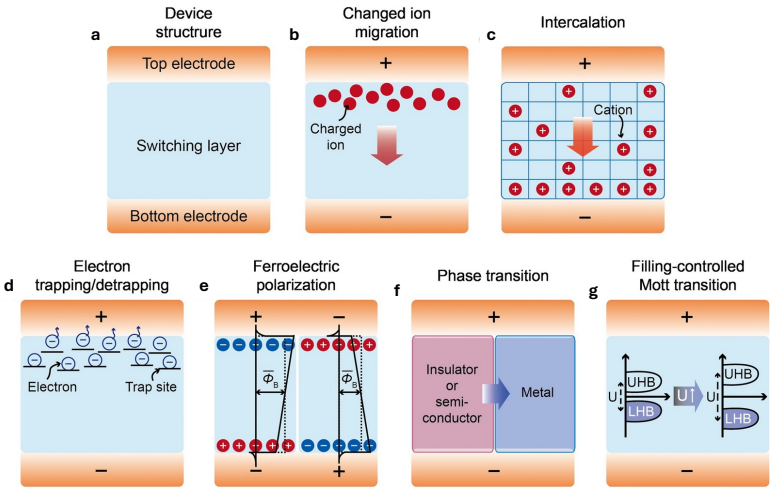


## 2.2. MEMRISTOR TECHNOLOGY FOR COMPUTING IN MEMORY

MEMRISTORS are two-terminal electronic devices that act as resistors with memory: their resistance can be modulated by the history of applied voltage or current, and they retain this state even when power is removed. First theorized by Chua in 1971 as the *fourth fundamental circuit element* [58], memristors remained hypothetical until 2008, when Strukov *et al.* at HP Labs reported the first physical implementation in a nanoscale $TiO_2$ device[59]. In essence, a memristor's conductance $G$ serves as an analog of a synaptic weight that can be increased or decreased by electrical stimuli, and once set, the conductance stays non-volatilely stored. This unique property, co-located storage and processing, makes memristors a natural building block for computing in memory . Memristors boast a simple structure (often a metal–insulator–metal sandwich), fast switching speeds, high density (nanoscale feature sizes), and compatibility with standard CMOS fabrication [59]. These traits have attracted significant industry interest in memristors for next-generation non-volatile memory and computing, with companies like Samsung, IBM, Intel, and others actively developing resistive memory (RRAM) technologies [59].

Several physical mechanisms can give rise to memristive behavior, leading to different device types. Oxide-based memristors (e.g., $TiO_2$, $HfO_2$) switch via the formation and dissolution of conductive filaments or changes in defect distributions under electric fields. Phase-change memory (PCM) devices (chalcogenides like GST) exhibit memristive behavior by reversible amorphous-crystalline phase transitions. Magnetic tunnel junctions can behave as memristors by changing resistance with magnetization (spintronic memristors). Ferroelectric memristors use polarization states. There are also organic memristors, 2D material memristors, and electrochemical metallization cells; each offers different switching speeds, endurance, ON/OFF ratios, etc.[59] Despite this variety, a common theme is the presence of multiple stable resistance levels that can be gradually programmed. This multi-level analog behavior is what allows memristors to act as artificial synapses: the conductance can be incrementally increased or decreased with voltage pulses, mimicking synaptic weight updates.

Memristors are especially relevant to neuromorphic computing because they




Figuur 2.1: Filament-free switching memristor structures and their switching mechanisms. (a) A typical simple two-terminal memristor structure consisting of electrode/switching layer/electrode. (b–g) Illustrations of the switching principles for diverse types of filament-free switching memristors. (b) Ion migration. The switching is mainly governed by the electric field-driven migration of the ions. (c) Intercalation. The electric field-driven intercalation of ions into lattices is responsible for the main switching behavior. (d) Electron trapping/detrapping. The electronic (de/)trapping processes at trap sites depending on the external electric field generates the switching events. (e) Ferroelectric polarization. The ferroelectric polarization switching results in the conductance switching. (f) Phase transition. A lower conductance phase (insulator or semiconductor) transforms into a higher conductance phase and vice versa. (g) Filling-controlled Mott transition. Mott–Hubbard gap increases or decreases corresponding to electron density in the transition metal orbitals. [57]

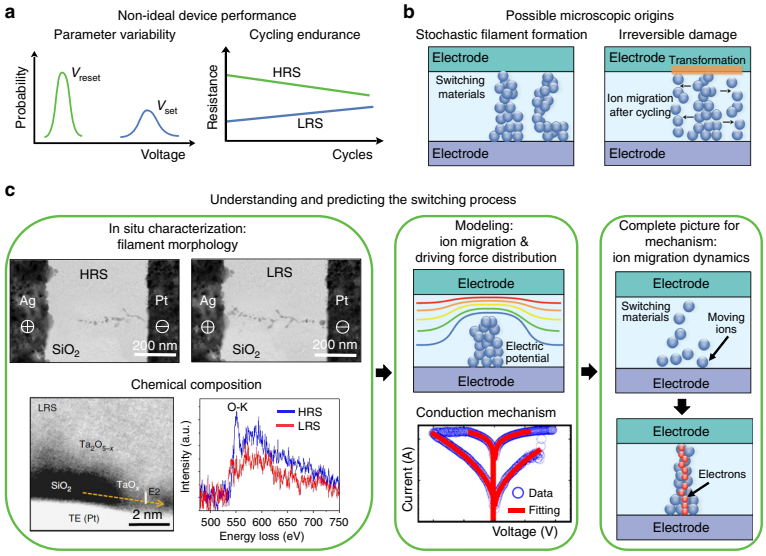
inherently realize the integrated memory-compute paradigm. Traditional CMOS implementations of synapses or memory elements might require dozens of transistors and capacitors per weight [59], whereas a single memristor can hold a weight in its conductance. Placed at the cross-points of a crossbar array of wordlines and bitlines, memristors naturally perform analog matrix-vector multiplication via Ohm's and Kirchhoff's laws: if input signals (e.g., voltages $V_i$) are applied to rows, the current summing at each column is

$$I_j = \sum_i G_{ij} V_i,$$

effectively computing a weighted sum in one physics step. This means an $M \times N$ memristor crossbar can compute

$$y = W x,$$

(with $W_{ij}$ encoded as conductances) in a single time slice, achieving massive parallelism. The result is a significant acceleration and energy reduction for




Figuur 2.2: Parameter variability of the set and reset voltages (left) and typical endurance failure behaviors in memristors (right). HRS represents the high resistance state, and LRS represents the low resistance state. b Possible microscopic origins responsible for the device behaviors: Conduction filaments in different forms contribute to the parameter variability (left of panel a); after cycling operation, the formation of an interfacial layer with high series resistance (orange disc) or the expansion of migration area for mobile ions (migrated blue balls marked by black arrows) leads to cycling endurance failure (right of panel a). c The in situ characterizations and device modeling can complement and complete each other and together provide a holistic picture of the resistive switching, as shown in the middle, where a device model is schematically presented with switching I–V loops of simulated and experimental data. Reproduced from ref. 12, Macmillan Publishers Ltd. (c), Ag/SiO_2Pt device; ref. 22, Macmillan Publishers Ltd. (c), Ta2O5-x/SiO2/Pt device. [57]

operations like neural network inference, which rely heavily on multiply-accumulate (MAC) computations. By eliminating the need to shuttle data between separate memory and processing units (the von Neumann “memory wall”), memristor-based computing in memory can overcome bottlenecks of conventional hardware.

Beyond speed, memristors exhibit device characteristics that emulate biological synapses. They can undergo synaptic plasticity updates in an analog manner; for instance, certain memristors have been shown to support spike-timing-dependent plasticity (STDP) when driven with appropriately shaped pulses, naturally implementing a form of learning rule. Their non-volatility means learned weights do not vanish when power is off, allowing neuromorphic systems to retain learned states without refresh. Memristors also operate at low voltages (a few volts or less) and can be extremely energy-efficient per operation (especially when exploiting

analog summation rather than digital logic toggling).

### 2.2.1. EMERGING MEMORY DEVICES



The behaviors of emerging memory devices are determined by their underlying physical mechanisms. These properties directly influence their suitability for neuromorphic computing systems, as some devices are optimized for nonvolatile analog resistive switching, while others with dynamic properties are better suited for constructing artificial neural units. In what follows, we list the main device types along with their working mechanisms, advantages, and drawbacks.

#### MEMRISTORS (RERAMS)

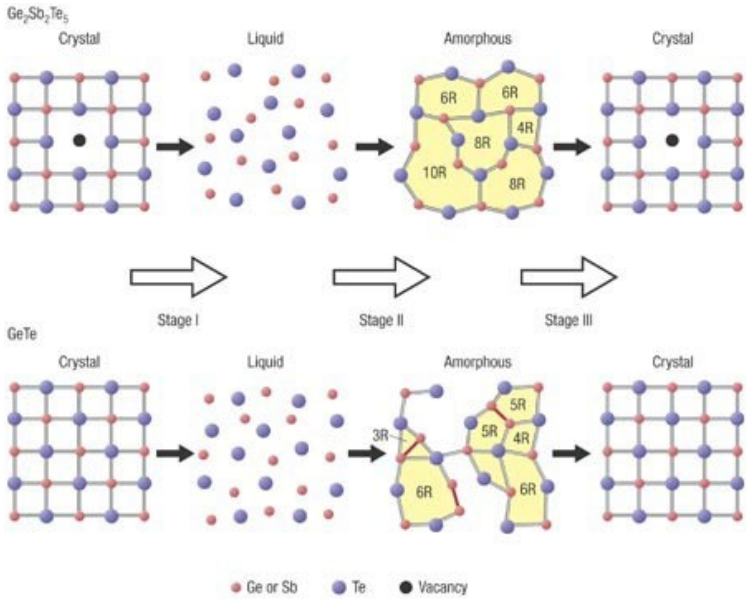
Memristors can be categorized into two primary types based on their switching mechanism. First, filament-type memristors can be further divided into cation filament and anion filament types [57, 60]. In cation filament devices, the switching is based on the migration and redox reactions of active metal ions, such as Ag or Cu [60]. In contrast, anion filament devices rely on the formation and rupture of conductive filaments due to the dynamics of oxygen vacancies. The second category comprises interfacial memristors, which operate by redistributing cations at the interface to modify the potential barrier and thereby alter the resistive state.

Memristors offer nonvolatile operation with nanosecond-level read latency and ultra-low write power in the picojoule to femtojoule range. They also feature high density and scalability, with device footprints in the order of tens of $F^2$, making them highly suitable for integration with CMOS technology. Moreover, their ability to support analog multilevel states enables them to represent finely tuned neural network weights.

However, memristors face several challenges. The stochastic nature of filament formation and dissolution may lead to significant cycle-to-cycle variability. In the case of interfacial devices, although the switching is smoother, the write speed is often slower and retention may be poorer. Additionally, achieving uniform behavior across a large array can be difficult due to device-to-device mismatch.

#### PHASE-CHANGE MEMORY (PCM)

PCM technology relies on the reversible phase transition of chalcogenide-based materials. The material can switch between a crystalline (low resistance) state and an amorphous (high resistance) state by controlled Joule heating. Crystallization is induced by maintaining a high temperature for a sufficient duration, whereas amorphization is achieved by rapid quenching, which locks the material into a disordered state [61], as illustrated in Fig. 2.3. PCM provides a nonvolatile memory solution with the ability to store multiple conductance states, allowing for stable analog representations [62]. On the downside, PCM typically requires high energy and longer pulse durations for the crystallization process, resulting in higher write energy and delay [63] as demenstrated in Fig. 2.4. Furthermore, the inherent phenomenon of resistance drift, where the resistance slowly changes over time, can compromise long-term accuracy [64, 65].




Figuur 2.3: The red bonds represent the Ge–Ge bond. Stage I and II: recording process in optical storage; stage III: erasing process in optical storage. Reprinted with permission from ref [66]

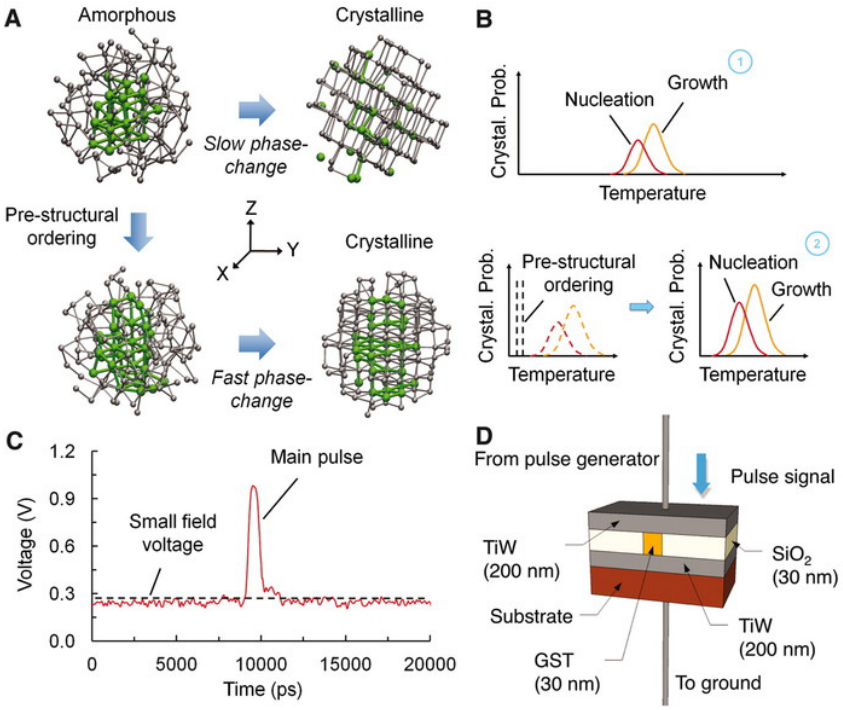
### FERROELECTRIC FIELD-EFFECT TRANSISTORS (FEFETS)

FeFETs are similar to conventional transistors but employ a ferroelectric insulator (commonly $HfO_x$) as the gate dielectric [67–69]. When a gate voltage is applied, the polarization in the ferroelectric layer switches, thereby modulating the conductance of the transistor channel. An accumulation effect is sometimes observed, which may result in abrupt switching after a series of pulses. One of the main benefits of FeFETs is their compatibility with CMOS fabrication processes, largely due to the use of $HfO_x$ [67]. They also exhibit good data retention and provide a moderate memory window, making them suitable for applications that require stable storage or synaptic plasticity emulation. The abrupt switching due to the accumulation effect can limit the granularity needed for smooth analog control, and FeFETs typically have a limited dynamic range compared to other analog devices. The recent research progress of FeFET are depicted in the Fig. 2.5

### ELECTROCHEMICAL RANDOM-ACCESS MEMORY (ECRAM)

ECRAM devices are three-terminal memory elements that include an insulating electrolyte layer between a metal gate and a conductive channel. By applying a gate voltage, ions are driven between the electrolyte and the channel, which modulates the device conductance in an analog manner. The deterministic nature of the ion motion contributes to a stable tuning of conductance.

A key advantage of ECRAM is its stable analog tuning ability, as the controlled ionic




Figuur 2.4: Prestructural ordering effects on the crystallization of PC materials. (A) Model configurations demonstrating the atomic rearrangements during the phase transition, with and without prestructural ordering. (B) Schematics of the crystallization probability as a function of temperature for PC materials. The nucleation and growth processes are accelerated when there is prestructural ordering (from method 1 to 2). (C) Waveform of a small-field incubation voltage and main pulse applied to set the PCRAM. A small voltage is first applied to initiate prestructural ordering, followed by a main pulse to induce crystal nucleation and growth of the PC material. (D) Schematic of the PCRAM structure with the pulse signal delivered to heat and crystallize the PC material (GST). Reprinted with permission from ref [63]

transport allows for reproducible conductance modulation. However, ECRAM faces integration challenges; issues such as ionic contamination and the high-temperature requirements of CMOS back-end processing pose significant hurdles. Moreover, concerns regarding the long-term chemical and mechanical stability of these devices persist and are still under investigation. Within the CiM (computing in memory ) framework, the crosspoint array physically represents the matrix of computational parameters (e.g., synaptic weights in a neural network) and performs MVM in the analog domain. As illustrated schematically in Fig. (a) (not shown here), applying a voltage $V_j$ at the $j$th column yields a current $I_i$ at the $i$th row (connected to

ground), given by:

$$I_i = \sum_{j=1}^{N} G_{i,j} \cdot V_j,$$

where $G_{i,j}$ denotes the conductance of the memory element at position $(i, j)$ and $N$ is the number of rows (or columns). In compact matrix-vector form, this relationship is written as:

$$\mathbf{i} = \mathbf{G} \cdot \mathbf{V}.$$

This mechanism underpins the physical execution of MVM in analog computing in memory systems, enabling highly parallel and energy-efficient computations that are critical for accelerating machine learning tasks.



### 2.2.2. CHALLENGES OF COMPUTING IN MEMORY (CIM)

Although analog computing in memory (CiM) offers low-power and high-throughput solutions for machine learning applications (especially for massive matrix–vector multiplications), several challenges remain:

#### ARCHITECTURAL LIMITATIONS

- CiM cores are typically optimized for MAC (multiply-accumulate) operations and hence best suited for applications that predominantly use these operations.
- Many large-scale machine learning algorithms also require activations, division, and other operations that are not ideal for current CiM architectures.
- CiM performs best in large arrays (e.g., $1024 \times 1024$), which can limit its use in smaller-scale applications.

#### APPLICATION-LEVEL CONSTRAINTS

- High-throughput inference is well-supported by CiM; however, during the training stage, limitations in endurance and write speed can become significant bottlenecks.
- Due to these constraints, CiM is more frequently targeted towards edge devices rather than centralized training environments requiring high computational capacity.

#### ACCURACY EVALUATION IN ANALOG COMPUTATION

- Input Error: Errors introduced by the digital-to-analog converters (DACs).
- Device Error: Variability in memory device conductance/resistance, where analog 8-bit precision is not equivalent to digital 8-bit precision.
- Array Error: Accuracy loss due to large-array effects such as IR-drop, crosstalk, etc.
- Output Error: Errors occurring in the analog-to-digital converters (ADCs).

Various integration concepts have been proposed:

- Conventional von Neumann Architecture: Separates memory and computing units on distinct chips, resulting in high latency and energy due to data movement.
- Near-Memory Computing (NMC): Embeds nonvolatile memory (eNVM) with the processing unit to reduce data movement.
- True CiM: Uses volatile SRAM or, ideally, integrates computation within the eNVM to further minimize data movement, though challenges in throughput, energy efficiency, and accuracy persist.

The co-design of device technologies, circuit engineering, and algorithms is essential to address these challenges and fully leverage the potential of emerging memory devices for CiM.

## 2.3. NEUROMORPHIC COMPUTING

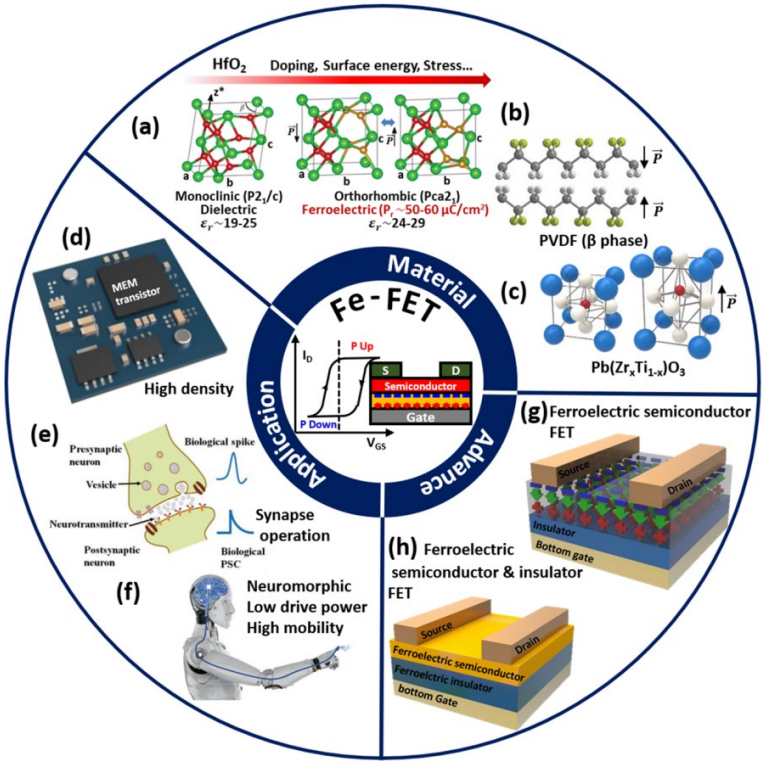
NEUROMORPHIC computing is a computational approach inspired by the structure, dynamics, and processing mechanisms of the human brain. It integrates concepts from neuroscience, computer science, and electrical engineering to develop systems that mimic neural networks, offering a path to more efficient, adaptive, and intelligent computing.

The motivations behind neuromorphic computing energy efficiency is that traditional computing architectures (von Neumann architecture) consume significant power for tasks like artificial intelligence (AI) and machine learning (ML), while neuromorphic systems aim to drastically reduce energy consumption by imitating the brain's energy-efficient signal processing. *Real-Time Adaptability:* Neuromorphic systems can process information in real time, making them suitable for dynamic environments where decision-making needs to be fast and adaptive. *Scalability and Parallelism:* Unlike traditional architectures, neuromorphic systems use highly parallel processing, similar to the brain, allowing them to handle large-scale, complex problems efficiently. *Cognitive Computing:* Neuromorphic computing targets human-like abilities such as learning, reasoning, and perception, which are challenging for classical AI models. *Biological Inspiration:* The brain's ability to process sensory information, adapt to new environments, and perform complex tasks with minimal energy motivates the development of systems that replicate these functions.

### 2.3.1. NEUROMORPHIC ARCHITECTURES: NON-SPIKING VS. SPIKING

Neuromorphic architectures are computing systems structured after the architecture of the brain, containing large numbers of "neurons" communicating via "synapses." They depart from the sequential, clocked operation of standard processors and instead emphasize parallel, event-driven computation. Two broad categories can be distinguished:

**Non-Spiking (Analog or Rate-Based) Neuromorphic Systems:** These systems implement neural networks using analog signals or digital circuits that do not rely on discrete spikes. Essentially, they realize conventional artificial neural networks (as




Figuur 2.5: Ferroelectric materials for the FeFET. (a) Hf0.5Zr0.5O2. Reproduced with permission from Park et al., J. Mater. Chem. 5, 4677 (2017). Copyright 2017 The Royal Society of Chemistry. (b) PVDF. Reproduced with permission from Chen et al., Adv. Electron. Mater. 3, 1600460 (2017). Copyright 2017 WILEY-VCH Verlag GmbH and Co.KGaA. (c) $Pb(Zr_xTi_{1-x})O_3$. FeFET applications. (d) MEM transistor capable of programming and erasing. (e) Synaptic transistor. Reproduced with permission from Chen et al., Adv. Electron. Mater. 6, 2000057 (2020). Copyright 2020 Weinheim, WILEY-VCH Verlag GmbH and Co.KGaA, Weinheim. (f) Realization of neuromorphic behavior through low power and high mobility. Reproduced with permission from Lee et al., Nat. Commun. 11, 2753 (2020). Copyright 2020 the Author(s) 2020, licensed under a Creative Commons Attribution 4.0 International License. Other structures using ferroelectric semiconductors. (g) Ferroelectric semiconductor-field effect transistor (FeS-FET). (h) FET using the ferroelectric semiconductor and insulator. Reprinted with permission from ref [70]

used in deep learning) in custom hardware with a brain-inspired layout. Examples include analog crossbar circuits with operational amplifiers or memristors, as well as mixed-signal accelerators. Here, neuron activations might be represented as continuous voltages or currents (or as digital numbers updated in parallel). The goal is to leverage physical parallelism and analog computation for speed and energy

gains without necessarily adopting the spike-based communication of biological neurons. Memristor crossbar arrays fall into this category, computing weighted sums in an analog fashion, while non-linear activation can be introduced using CMOS neurons at the outputs.



**Spiking Neuromorphic Systems:** These architectures more directly emulate biological information processing by using spikes, discrete, binary pulses fired by neurons when their membrane potential crosses a threshold. Spiking neural networks (SNNs) communicate via asynchronous spike events rather than continuous signals. Hardware implementations of SNNs include digital chips (such as IBM's TrueNorth and Intel's Loihi) and mixed analog/digital designs (e.g., BrainScaleS and DYNAP chips) that implement neuron circuits (commonly leaky integrate-and-fire models) and synapse circuits (which can be memristive devices or capacitive weights). In spiking systems, information is encoded in spike rates or timing patterns. A key feature is that computation is event-driven: if there are no spikes, the system remains largely idle, leading to extremely low power usage for sparse activity. This event-driven property is thought to be a major factor behind the human brain's efficiency (approximately 20 W power consumption). Spiking systems naturally support low-power and real-time processing, making them particularly attractive for applications such as adaptive tomography or error diagnostics in quantum processors.

## 2.4. SPIKING NEURAL NETWORKS WITH MEMRISTIVE DEVICES

SPIKING Neural Networks (SNNs), known as third-generation neural networks, mimic the spiking behavior of biological neurons, processing information via discrete spikes rather than continuous signals [71]. The spike-timing-dependent plasticity (STDP) learning rule, a biologically inspired method adjusting synaptic weights based on timing differences between spikes, is integral to SNN training [71]. Memristive devices, particularly Phase Change Memory (PCM) and ReRAM, have demonstrated significant promise in implementing SNNs due to their tunable resistance characteristics. Kuzum et al. (2012) first demonstrated STDP with PCM devices (Fig. 2.2(a)) [13], while subsequent research explored various synaptic behaviors and plasticity with both PCM and RRAM (Fig. 2.2(b)). Additional memristive technologies, including magnetic and ferroelectric tunnel junctions, further extend the device possibilities for efficient SNN implementations [72, 73].

### 2.4.1. NEURAL NETWORK MODELS FOR QUANTUM STATE TOMOGRAPHY

A variety of neural network models have been explored for quantum state tomography, each with different representational capacities and training requirements. Below we briefly review the main approaches:

Restricted Boltzmann Machines (RBMs): RBMs are two-layer generative models with a visible layer (representing measurement outcomes) and a hidden layer, with binary stochastic neurons and no intra-layer connections. Carleo and Troyer

introduced Neural Quantum States with RBMs to encode wavefunctions, capturing complex entanglement with relatively few parameters. Torlai *et al.* later demonstrated QST using RBMs to learn states from projective measurements. RBMs are compatible with neuromorphic hardware due to their simple, bipartite structure and can be efficiently mapped onto memristor crossbar arrays.

Convolutional Neural Networks (CNNs): Widely used in image processing, CNNs have been adapted for QST, particularly for systems exhibiting spatial structure. A CNN architecture leverages convolutional layers to impose translational invariance and capture local correlations, thereby reducing the number of parameters needed. Schmale *et al.* introduced a CNN-based QST scheme that achieved high fidelity using a variational ansatz with polynomial scaling in system size. Although convolution operations entail challenges such as weight sharing, they can be implemented via time-sequential methods or specialized hardware designs.

Recurrent Neural Networks (RNNs): RNNs, including gated variants like LSTMs and GRUs, are adept at handling sequential data. In QST, RNNs can be used to sequentially model the probability distributions of measurement outcomes, processing data qubit by qubit or adaptively updating state estimates. Huang *et al.* used bidirectional GRUs to reconstruct both separable and highly entangled states with fewer measurements. Neuromorphic implementations may leverage natural recurrent dynamics of spiking networks for such sequential processing.

Generative Adversarial Networks (GANs): GANs consist of a generator that creates candidate density matrices and a discriminator that distinguishes generated data from real measurements. Conditional GANs (CGANs) have been applied to QST, achieving high-fidelity reconstructions with substantially less data than conventional methods. While training GANs on neuromorphic hardware presents challenges, inference with pre-trained networks is feasible for rapid state estimation.

Transformer Networks: Transformers use self-attention mechanisms to model long-range correlations. In QST, transformer-based models have been developed to capture global entanglement across qubits effectively. Cha *et al.* presented an attention-based quantum tomography method that outperformed earlier RNN-based approaches. Although the computational complexity of attention operations poses implementation challenges in neuromorphic hardware, ongoing research aims to devise hardware-friendly approximations.

Variational Autoencoders (VAEs): VAEs consist of an encoder and a decoder that map input measurement distributions to a low-dimensional latent space and reconstruct the quantum state from that latent representation. This approach naturally handles noise and uncertainty. Liang *et al.* demonstrated that a VAE could reconstruct GHZ states with high fidelity from measurement data. The decoder, being feed-forward, is particularly well-suited to neuromorphic implementation.

### 2.4.2. CLASSICAL VS. NEUROMORPHIC IMPLEMENTATION: SCALABILITY, ENERGY, AND ACCURACY

**Scalability:** Classical computing systems (CPUs/GPUs) can simulate neural networks of arbitrary size, but simulation time and memory usage grow rapidly. Neuromorphic hardware, by physically replicating neuron and synapse circuits, avoids the von

Neumann bottleneck, enabling parallel computation even in very large networks. For example, Intel's Loihi-based systems have demonstrated real-time performance with networks comprising millions of neurons. Memristor crossbar accelerators can perform $O(N^2)$ operations in constant time, offering promising scalability for QST as network size increases.

Energy Efficiency: Neuromorphic systems achieve extraordinary energy efficiency by integrating memory with processing and utilizing event-driven computation. Reports indicate that spiking neuromorphic systems (e.g., Loihi 2) may consume up to 100× less energy than conventional CPUs/GPUs, while operating 50× faster Analog memristor crossbars can reach energy per MAC operation in the femtojoule range, far lower than digital logic. This efficiency is crucial when processing large datasets or performing iterative neural network evaluations for QST.

Accuracy: Classical implementations benefit from high-precision computations (typically 32-bit floating point), ensuring accurate state reconstruction. Neuromorphic hardware, especially analog and spiking systems, introduces noise, quantization errors, and device variability. Nonetheless, many studies have shown that with proper calibration and noise-aware training techniques, neuromorphic systems can achieve accuracy comparable to classical hardware. Hybrid approaches may leverage neuromorphic systems for fast, initial state estimates with subsequent refinement on classical processors if needed.

### 2.4.3. LIMITATIONS AND FUTURE DIRECTIONS

Despite significant promise, several challenges remain:

- Device-Level Limitations: Variability in memristor switching, resistance drift, and limited endurance can affect neural network performance. Ongoing research aims to improve device materials, incorporate error-correcting schemes, and develop multi-device synapses to mitigate these issues.
- Training and Algorithmic Challenges: Implementing full backpropagation on neuromorphic hardware is challenging. Research into in-situ training using local rules (e.g., STDP) or hybrid quantum-classical training schemes is ongoing.
- Integration with Quantum Systems: Interfacing neuromorphic hardware with high-rate quantum measurement setups remains a technical hurdle. Future work may focus on minimizing latency through co-location or cryogenic integration.
- Accuracy and Theoretical Guarantees: Incorporating physical constraints (e.g., complete positivity) into neural network models and establishing robust error bounds is critical for ensuring reliable state estimation.
- Interdisciplinary Co-Design: Collaboration between quantum physicists, machine learning researchers, and hardware engineers is essential to tailor neural network architectures to the strengths and limitations of neuromorphic substrates.

Addressing these challenges will be key to deploying neuromorphic hardware as a standard tool in quantum state tomography, enabling real-time, efficient, and scalable quantum information processing.



### CHALLENGES IN NEUROMORPHIC COMPUTING

Hardware Limitations: Current electronic hardware struggles to replicate the complexity and density of biological neural networks. The design of memristors, synaptic devices, and spiking neural networks (SNNs) is still in its infancy.

Algorithm Development: Traditional algorithms used in AI are not directly compatible with neuromorphic architectures. New algorithms optimized for spiking neural networks and asynchronous processing are required. Building large-scale neuromorphic systems that maintain low power consumption while scaling to billions of neurons and synapses remains a significant challenge (Scalability Issues). Neuromorphic systems use event-based data representation, unlike conventional systems that process data in a deterministic, step-wise manner. This shift demands new paradigms in data handling and interpretation. (Data Representation). Neuromorphic systems need to integrate with existing digital and classical computing platforms, creating compatibility challenges. (Integration with Existing Systems). There is no universal standard for neuromorphic architectures, leading to fragmentation in development and difficulty in comparing approaches. (Lack of Standards).

Problems to Address: Neuromorphic devices rely on emerging materials like phase-change materials or memristive technologies, which are not yet fully optimized for mass production. (Material Challenges). Biological systems tolerate noise and errors well, but replicating this robustness in neuromorphic hardware remains difficult. (Noise and Reliability). The human brain's functioning is not fully understood, limiting the ability to create accurate models and architectures. (Lack of Universal Models). Unlike conventional neural networks, neuromorphic systems need event-driven and unsupervised learning algorithms, which are less mature and harder to train. (Training and Learning Paradigms). Neuromorphic systems are expensive to develop and manufacture, creating barriers for widespread adoption in industries. (Economic Viability).

Future Directions: Advancing Memristive Devices: Research is focused on developing reliable, scalable, and low-power memristors to act as artificial synapses in neuromorphic systems. Brain-Inspired Learning Algorithms: Developing biologically plausible learning mechanisms, such as spike-timing-dependent plasticity (STDP), to improve adaptability. Hybrid Architectures: Combining traditional computing with neuromorphic systems to leverage the strengths of both paradigms. Applications in AI and Robotics: Neuromorphic systems have potential applications in autonomous systems, real-time sensory processing, and adaptive robotics. Neuromorphic computing represents a revolutionary step toward energy-efficient, intelligent systems capable of mimicking human-like cognition. However, realizing its full potential requires overcoming significant technical, algorithmic, and economic challenges.

## 2.5. CONCLUSION OF THIS CHAPTER

MACHINE learning and neuromorphic hardware are poised to transform quantum state tomography by addressing the exponential complexity of high-dimensional quantum systems. Neural network models can learn to represent quantum states with drastically fewer parameters than traditional methods, while memristor-based neuromorphic systems offer unparalleled parallelism and energy efficiency. The literature reveals successful demonstrations of RBMs, CNNs, RNNs, GANs, Transformers, and VAEs for QST, with various approaches showing promising scalability and speed improvements. Although challenges such as device variability, training limitations, and system integration remain, ongoing advances suggest that neuromorphic hardware will play a crucial role in future quantum state estimation. By leveraging interdisciplinary innovation, the convergence of quantum physics, machine learning, and specialized hardware holds the promise of enabling real-time tomography and broader quantum-classical hybrid computing architectures. The next chapter formalizes these ideas by providing the theoretical background, material physics and QST mathematics, that support subsequent device implementation.

# 3

# THEORETICAL BACKGROUND

## 3.1. LIQUID-SI FOR SILICON DEVICES

A key enabler for integrating neuromorphic circuits on unconventional substrates is the use of liquid-phase silicon precursors in conjunction with electron-beam processing. Liquid silicon (L-Si) precursors, such as polysilanes (long-chain silicon-based polymers), can be *spin-coated* onto a substrate to form a thin film. Subsequently, a focused electron beam is used to convert this liquid-deposited film into solid silicon via **electron-beam-induced processing (EBIP)**. In contrast to traditional chemical vapor deposition or furnace annealing, EBIP can be performed at room temperature, avoiding high thermal budgets. The electron beam locally dissociates and restructures the polysilane material, leaving behind a silicon network or film. This process effectively "writes" silicon structures with *nanoscale resolution* while operating at ambient temperature. As a result, EBIP eliminates the need for high-temperature annealing to crystallize or solidify the silicon, which is advantageous for substrates that cannot tolerate high heat (e.g. plastics or pre-fabricated circuits).

The advantages of Liquid-Si EBIP are multifold. First, processing at or near room temperature dramatically reduces energy consumption and minimizes thermal stress and inter-diffusion in multi-layer stacks. It also mitigates contamination since the deposition and conversion can occur in a controlled vacuum or inert environment, yielding high-purity silicon structures. Second, EBIP's direct-write nature grants exceptional patterning precision. Feature sizes on the order of tens of nanometers are achievable, allowing fine control over device dimensions and nanoscale architectures. This precision is critical for optimizing electronic properties (e.g. reducing leakage currents in nanodevices) and for exploring novel device physics at the nanoscale. Third, the combination of spin-coating and EBIP is compatible with a wide range of substrates. Since no high-temperature step is required, one can fabricate silicon devices on flexible polymer foils, glass, or even paper, in addition to conventional silicon or oxide wafers. This opens possibilities for **flexible electronics** and heterogeneous integration where neuromorphic circuits could be embedded on sensors or directly on quantum hardware components without damaging them.

Notably, Shimoda demonstrated as early as 2006 that solution-processed silicon films could be used to create transistors on glass, heralding the concept of liquid silicon electronics. Recent progress builds on this by using EBIP as a novel means to convert solution-deposited silicon precursors into device layers at low temperature. The successful fabrication of high-quality silicon nanopatterns using EBIP has been reported, indicating that the material produced can have sufficient electronic quality for device operation. In the context of neuromorphic hardware, this Liquid-Si EBIP platform means that we can integrate silicon-based memristive devices or neuron transistors on atypical substrates or in advanced 3D architectures that were previously infeasible. It paves the way for monolithic integration of silicon neuromorphic components with other systems (for example, on the same chip as a quantum processor or photonic system), since the entire process is CMOS-friendly and additive. In summary, Liquid-Si EBIP provides a materials foundation for building silicon neuromorphic devices with high resolution and low thermal budget. This will be leveraged in our hybrid system to fabricate dense crossbar arrays of

synaptic elements on chip, as well as interface circuitry, without compromising the delicate quantum components.

## 3.2. RESISTIVE RAM (RERAM) PHYSICS



NEUROMORPHIC computing architectures often rely on **memristive** devices to act as artificial synapses that store and update connection weights. Among various memristive technologies, resistive random-access memory (ReRAM) devices have gained prominence due to their scalability, non-volatility, and analog resistive switching behavior. In a ReRAM (also called *resistive switching memory*), an insulating or semiconducting dielectric layer is sandwiched between two electrodes. By applying appropriate voltage pulses, the device can be switched between a high-resistance state (HRS, "OFF") and a low-resistance state (LRS, "ON"). This switching is governed by the formation or dissolution of conductive paths within the dielectric, a phenomenon known as **resistive switching**. The physical mechanisms for resistive switching can vary, but two prevalent categories are filamentary conduction via ionic migration and interface/bulk charge trapping mechanisms.

In many oxide-based ReRAMs (such as $HfO_2$, $TiO_2$, $Ta_2O_5$, etc.), the switching is filamentary and driven by the migration of oxygen anions (or equivalently, oxygen vacancy diffusion). This is often referred to as a *valence change mechanism* (VCM). When a sufficient electric field is applied, oxygen anions are drawn out of a region in the oxide towards one electrode, leaving behind a nanoscale filament of oxygen vacancies that is more conducting (essentially a reduced oxide). This filament connects the two electrodes, setting the device to LRS. A reverse voltage pushes oxygen back into the filament, disrupting the conductive path and resetting the device to HRS. These conductive filaments can be visualized in electron microscopy; for instance, in $TiO_2$ ReRAM devices a Magnéli-phase filament ($Ti_4O_7$) forms and can be observed post-switching. The filamentary model explains the abrupt resistance changes and the memory effect (the filament remains until intentionally broken). Another class of ReRAM uses cation migration, an *electrochemical metallization (ECM)* mechanism, where metal ions (from an active electrode like Ag or Cu) drift into the dielectric and plate out a metallic filament. This is sometimes called conductive-bridge RAM (CBRAM). ECM-based devices similarly toggle between insulating and metallic states, but often at very low forming voltages, albeit sometimes with lower retention.

Regardless of mechanism, a common requirement in many ReRAM devices is an initial **electroforming** step. Electroforming is a one-time high-voltage pulse that creates a permanent seed filament or conductive channel in a fresh device that is initially insulating. This step introduces the first conductive path but can be stochastic and cause device-to-device variability as well as potential damage. Forming generally increases the device's leakage and can limit yield and endurance. For large-scale integration of neuromorphic systems, *forming-free* ReRAM is highly desirable. Researchers have found that by engineering the materials stack, forming can be avoided. For example, using an oxide that is slightly oxygen-deficient or adding a layer that acts as an oxygen vacancy reservoir can pre-form conductive

pathways. In particular, **hafnium oxide ($HfO_2$)** has emerged as a promising dielectric for forming-free operation due to its high dielectric strength and compatibility with CMOS processes. When $HfO_2$ is combined with certain oxygen-scavenging electrodes (such as palladium), an interfacial oxide layer with pre-created oxygen vacancies can form (e.g. Pd-O-Hf interface) that lowers the barrier for filament formation. This means the device can switch without a high-voltage forming step, achieving *forming-free behavior*. Such $HfO_2$-based ReRAM devices not only avoid the energy and variability costs of forming, but also tend to exhibit lower switching voltages and more reproducible multi-level states. Indeed, engineering the electrode materials and oxide stoichiometry has led to ReRAM cells that switch reliably out-of-the-box and support multiple resistive levels (instead of just binary).

Analog switching and synaptic weight storage: Unlike digital flash memory which has two states, memristive ReRAM devices can often be programmed to intermediate resistance values. By controlling the magnitude or duration of the programming pulses, one can achieve a continuum (or at least multiple discrete levels) of conductance. This behavior is crucial for emulating biological synapses, which strengthen or weaken in analog fashion. The ReRAM's conductance $G$ can represent a synaptic weight $w$ in a neural network. When a voltage corresponding to a neuron's spike or activation is applied, the current through the memristor ($I = GV$) naturally implements a weighted contribution. Neuromorphic systems take advantage of this by arranging memristors in crossbar arrays where Kirchhoff's circuit laws perform vector-matrix multiplication in one step: if an array has conductances $G_{ij}$ at crosspoint $(i, j)$, and a vector of input voltages $V_j$ is applied to each column $j$, the current summing into row $i$ is

$$I_i = \sum_j G_{ij} V_j$$

which is exactly the dot-product of the input vector with the $i$th row of the conductance matrix. This corresponds to a weighted sum in a neural network layer. By programming the ReRAM conductances to appropriate values, the crossbar performs inference or computation in a single analog step, achieving massive parallelism.

Because of their filamentary nature, real ReRAM devices exhibit some non-idealities from a neural network perspective: device-to-device variability, cycle-to-cycle drift, and nonlinear weight update dynamics. However, extensive research has shown that by tailoring pulse schemes or using arrays of devices per weight, these issues can be mitigated. ReRAM-based synapses have successfully demonstrated key learning rules such as spike-timing-dependent plasticity (STDP) by appropriate timing of pulses to modulate conductance. Additionally, short-term plasticity effects (facilitation and depression) can emerge from volatile resistive switching in certain devices, providing a closer emulation of real synapses. From a modeling perspective, the ReRAM can be described by internal state variables (e.g. filament length or vacancy density) that evolve according to differential equations when voltage is applied. One simple phenomenological model is

$$\frac{dx}{dt} = \eta f(x) I(t), \tag{3.1}$$

where:

- $x \in [0,1]$ represents the state of the device, with $x = 0$ corresponding to the high resistance state (HRS) and $x = 1$ corresponding to the low resistance state (LRS).
- $\eta$ is a material-dependent mobility or rate factor.
- $f(x)$ encapsulates the voltage threshold and limiting behavior. For example, one might have $f(x) \approx 0$ when $x$ is near 0 or 1, mimicking the fact that the device stops changing significantly when it is fully off or fully on.
- $I(t)$ is the applied current as a function of time.



The instantaneous resistance $R(x)$ or conductance $G(x)$ is a deterministic function of $x$. Such models, including the TEAM/VTEAM models, allow circuit simulators to incorporate memristive synapses and predict their analog computing performance.

For our hybrid neuromorphic-quantum system, ReRAM devices serve as the core memory elements that store the parameters (weights) of a neural network designed to perform quantum state tomography. The ReRAM crossbar can execute linear computations (matrix-vector products) extremely efficiently in hardware, which is essential given the potential size of the tomography problem for large quantum states. Furthermore, the forming-free $HfO_2$ devices we employ offer stable, repeatable multi-bit operation, meaning they can represent the finely tuned probabilities or amplitudes needed for high-fidelity reconstruction. By leveraging these analog memory properties, we aim to implement on-chip learning rules that update synaptic weights based on incoming data (e.g. measurement results from a quantum experiment) in real time. The energy efficiency of such an approach is noteworthy: computations occur *in situ* in the memory, drastically reducing data movement. This can lead to orders-of-magnitude lower power consumption compared to conventional digital processors.

In sum, ReRAM physics provides a mechanism to emulate synapses *in silico*, and its modeling informs how we program and use these devices in a neuromorphic computing context. With this solid-state synaptic array in hand, we now turn to the theory of quantum state tomography to see how these "silicon synapses" can be used to reconstruct quantum states. Understanding these conduction and noise mechanisms is essential for predicting cryogenic stability, as analyzed in Chapter 6

## 3.3. QUANTUM STATE TOMOGRAPHY: FORMALISM AND RECONSTRUCTION

QUANTUM state tomography (QST) is the process of determining the full quantum state of a system (represented by a density operator $\rho$) from measurement data. In quantum mechanics, unlike in classical measurement, observing a system generally disturbs it, and a single measurement yields only partial information (one outcome) about the state. Therefore, to reconstruct $\rho$ one must prepare many

identical copies of the quantum system and measure each copy, often using a variety of measurement settings.

By gathering the statistical outcomes of these measurements, one can infer the state. Tomography is essentially an *inverse problem*: given the measured data, find the $\rho$ that most likely produced it. Formally, a quantum state in a $d$-dimensional Hilbert space is described by a $d \times d$ density matrix $\rho$ which is positive semi-definite (all eigenvalues $\geq 0$) and has unit trace ($\mathrm{Tr}[\rho] = 1$). For a state of $n$ qubits, $d = 2^n$, so $\rho$ is $2^n \times 2^n$ and has $d^2 - 1$ independent real parameters. The exponential growth of parameters with $n$ underscores the challenge of QST.



In a typical QST experiment, one performs a set of measurements on the quantum state. Each distinct measurement is characterized by a set of outcomes with known operators (more formally, a Positive Operator-Valued Measure (POVM) as discussed in the next section). For example, for qubits one might choose to measure along the $X$, $Y$, and $Z$ axes of the Bloch sphere. Each measurement setting yields outcome counts that can be converted into empirical probabilities. The relationship between the state and measurement outcomes is given by **Born's rule**:

$$p(i) = \mathrm{Tr}(\rho E_i),$$

where $\{E_i\}$ are the measurement operators with $E_i$ corresponding to outcome $i$. By collecting enough different outcome probabilities $p(i)$ from a sufficiently informative set of measurements, one can attempt to solve for the unknown $\rho$. In practice, the measurements are repeated many times to estimate frequencies $f_i \approx p(i)$, since quantum probabilities are accessed via relative frequencies.

A simple approach to QST is **linear inversion**. If one has as many independent equations as the number of unknowns in $\rho$, one can invert the linear system to find an estimate $\hat{\rho}_{\mathrm{lin}}$. For instance, in single-qubit tomography, measuring the Pauli $X$, $Y$, and $Z$ expectation values provides enough information to reconstruct the Bloch vector (and hence $\rho$). However, linear inversion often yields an estimator that is not physical; due to statistical noise, the reconstructed matrix may not be positive semidefinite (it might have small negative eigenvalues). This is undesirable since a valid quantum state cannot have negative probabilities.

To address this, the standard method in QST is **maximum-likelihood estimation (MLE)**. In MLE, one seeks the density matrix $\hat{\rho}$ that maximizes the likelihood of the observed data (or equivalently minimizes the discrepancy between predicted and observed frequencies) while imposing the constraints $\rho \geq 0$ and $\mathrm{Tr}[\rho] = 1$. Assuming independent trials, the likelihood function given data $\{n_i\}$ (counts for outcome $i$) is

$$L(\rho) = \prod_i \left[\mathrm{Tr}(\rho E_i)\right]^{n_i},$$

and one typically maximizes the log-likelihood

$$\ln L(\rho) = \sum_i n_i \ln \mathrm{Tr}(\rho E_i),$$

often using iterative algorithms such as gradient ascent or expectation-maximization.

The result $\hat{\rho}_{\mathrm{MLE}}$ is guaranteed to be a valid density matrix by construction of the optimization (usually one parameterizes $\rho = \frac{T^\dagger T}{\mathrm{Tr}(T^\dagger T)}$ for some lower-triangular $T$ to enforce positivity).

James *et al.* provide a comprehensive description of qubit state tomography using both linear inversion and MLE, showing that MLE yields physical states and discussing error bars for the estimates.



Informational completeness and efficiency: The choice of measurements in QST must be such that the data is *informationally complete* – i.e., the set of measured observables spans the space of all possible states. In practical terms, one needs at least $d^2 - 1$ independent probability constraints to solve for $d^2 - 1$ unknowns (for a $d$-dimensional density matrix). A common tomographic strategy is to measure a basis of observables. For example, for a single qubit, measuring $\sigma_X$, $\sigma_Y$, $\sigma_Z$ (each providing one independent expectation value) suffices. For two qubits, one could measure tensor products of Pauli matrices (9 settings) to get a tomographically complete set. In general, for $n$ qubits one might choose $3^n$ Pauli-product measurements to be complete. There are also more efficient designs: a **symmetric informationally-complete POVM (SIC-POVM)** uses $d^2$ outcomes (for a $d$-dimensional system) in a single generalized measurement to achieve informational completeness with minimal redundancy. Such measurements have each $E_i$ of equal norm and equal pairwise angles, and have been studied theoretically as an optimal solution to minimize statistical error, though exact SIC-POVM constructions are known only for certain dimensions.

The number of measurements needed and the post-processing overhead are the major bottlenecks of conventional QST. The number of parameters grows exponentially with system size, which means full tomography becomes impractical for even moderately large $n$ (for $n > 10$ qubits, $d^2 > 10^6$ parameters). Indeed, brute-force tomography demands exponential resources and is infeasible beyond small systems. This has motivated a variety of more efficient approaches such as compressed sensing QST (assuming the state is low-rank or has structure and adaptive tomography (where measurement settings are chosen on-the-fly to maximize information gain). Nevertheless, a recent direction is to leverage **machine learning** to handle the complexity of QST. Machine learning methods, especially neural networks, can offer a clever parameterization of quantum states or an efficient way to process tomographic data, thereby drastically reducing the effective number of parameters or accelerating the reconstruction algorithm.

In the following sections, we discuss the general framework of POVMs for quantum measurements and then delve into how neural networks,in particular Spiking Variational Autoencoders (SVAE), can be used for quantum state reconstruction, forming the “brain” of our hybrid neuromorphic-quantum system. The statistical reconstruction discussed here provides the theoretical basis for the machine-learning inference in Section 3.5.

## 3.4. SYMMETRIC INFORMATIONALLY COMPLETE POVMS (SIC-POVMS)

SYMMETRIC Informationally Complete Positive Operator-Valued Measures (SIC-POVMs) represent a class of measurement schemes that are informationally complete and possess a high degree of symmetry. In the context of QST, SIC-POVMs are particularly valuable because they offer an optimal balance between measurement efficiency and information extraction, making them ideal for scalable implementations.



### 3.4.1. POVM FORMALISM

In the standard quantum measurement framework, measurement outcomes are associated with a set of operators $\{M_a\}$ forming a POVM, which satisfies:

$$M_a \geq 0, \quad \sum_a M_a = \mathbb{I}. \tag{3.2}$$

Each operator $M_a$ corresponds to a possible measurement outcome $a$, and the probability of observing outcome $a$ for a given state $\rho$ is given by the Born rule:

$$P(a|\rho) = \mathrm{Tr}(\rho M_a). \tag{3.3}$$

A POVM is said to be *informationally complete* if the measurement statistics $\{P(a|\rho)\}$ uniquely determine the state $\rho$. Among these, SIC-POVMs are distinguished by maximal symmetry and minimal number of outcomes.

### 3.4.2. DEFINITION OF SIC-POVM FOR SINGLE QUBIT

For a single qubit, the SIC-POVM consists of four rank-one projectors $\{|\psi_i\rangle\langle\psi_i|\}_{i=0}^{3}$, scaled to satisfy the completeness relation:

$$M_i = \frac{1}{4}|\psi_i\rangle\langle\psi_i|, \quad \sum_{i=0}^{3} M_i = \mathbb{I}. \tag{3.4}$$

The four SIC-POVM states $\{|\psi_i\rangle\}$ form the vertices of a regular tetrahedron inscribed in the Bloch sphere. These states are:

$$|\psi_0\rangle = |0\rangle, \tag{3.5}$$

$$|\psi_1\rangle = \frac{1}{\sqrt{3}}|0\rangle + \sqrt{\frac{2}{3}}|1\rangle, \tag{3.6}$$

$$|\psi_2\rangle = \frac{1}{\sqrt{3}}|0\rangle + \sqrt{\frac{2}{3}}e^{i\frac{2\pi}{3}}|1\rangle, \tag{3.7}$$

$$|\psi_3\rangle = \frac{1}{\sqrt{3}}|0\rangle + \sqrt{\frac{2}{3}}e^{-i\frac{2\pi}{3}}|1\rangle. \tag{3.8}$$

These states are equiangular, satisfying:

$$|\langle\psi_i|\psi_j\rangle|^2 = \frac{1}{3} \quad \forall i \neq j. \tag{3.9}$$

### 3.4.3. CONSTRUCTING POVMS FOR MULTI-QUBIT SYSTEMS

For an $n$-qubit system, the total SIC-POVM is formed by taking tensor products of single-qubit SIC-POVM elements. For example, for two qubits:

$$M_{a_1 a_2} = M_{a_1} \otimes M_{a_2}, \quad a_1, a_2 \in \{0,1,2,3\}. \tag{3.10}$$

This results in $4^2 = 16$ distinct measurement operators for a two-qubit system. In general, an $n$-qubit system will yield $4^n$ measurement outcomes.

### 3.4.4. APPLICATION IN QST

The SIC-POVM framework is ideal for quantum state tomography due to:

- Minimal Measurement Set: Only $d^2$ outcomes are required for a $d$-dimensional system, the minimal number for full state reconstruction.
- Uniform Treatment: All POVM elements are symmetric, avoiding bias in the measurement scheme.
- Efficiency in Learning: In the context of SVAE, the one-hot encoded measurement outcomes obtained via SIC-POVMs serve as ideal input for variational models, enabling fast convergence and high-fidelity reconstructions.

### 3.4.5. EXAMPLE: TWO-QUBIT SIC-POVM

Let $M_i^{(1)}$ be the SIC-POVM elements for qubit 1 and $M_j^{(2)}$ for qubit 2. Then:

$$M_{ij} = M_i^{(1)} \otimes M_j^{(2)}, \quad i, j = 0,1,2,3. \tag{3.11}$$

This gives a complete basis of 16 linearly independent POVM elements for two qubits. The measured probabilities $P_{ij}$ from repeated measurements are then used to reconstruct the quantum state via linear inversion or learned inference in models like SVAE.

## 3.5. NEURAL NETWORK AS AN INFERENCE TOOL

IN a different vein, one can train a neural network to directly predict the parameters of $\rho$ from raw measurement data. This supervised learning approach trains the network on many examples of synthetic data paired with the true state, so that it learns an inverse mapping. Once trained, the network can quickly output an estimated density matrix from new experimental frequencies. However, this method requires a representative training set of quantum states, which is challenging given the vast state space. It may be feasible if one restricts the approach to certain classes

of states (for example, low-rank states or those parameterized by a quantum circuit of limited depth).

In practice, unsupervised or semi-supervised methods are often preferred, where the network is trained on a single dataset (similar to training an RBM or variational autoencoder on measured data). Neural networks can also be used to adaptively choose measurements (an active learning strategy for QST), or to denoise reconstructed states (e.g. using neural nets to map a noisy density matrix to a physically plausible one).



Mathematically, training a generative neural network for QST involves defining a loss function based on the data likelihood or on a divergence between the network's predicted distribution $q_\theta(i)$ and the actual distribution $p(i)$. For example, one may minimize the Kullback–Leibler (KL) divergence

$$D_{\mathrm{KL}}\big(\{f_i\} \,\|\, \{q_\theta(i)\}\big) = \sum_i f_i \ln \frac{f_i}{q_\theta(i)},$$

where $f_i$ is the observed frequency of outcome $i$. Minimizing this divergence corresponds to maximizing the likelihood that the network's distribution generated the observed data. Constraints like $\sum_i q_\theta(i) = 1$ are automatically satisfied by, for example, using a softmax output layer.

To ensure the positivity of the reconstructed $\rho$ when using a neural network ansatz, one may parameterize $\rho$ as $A^\dagger A$ (which is always positive) or output elements corresponding to the Cholesky decomposition of $\rho$. In RBMs or similar models, contrastive divergence or other training techniques are commonly employed. In essence, the neural network provides a differentiable map from latent parameters to outcome probabilities, with optimization ensuring agreement with the data.

From a representational standpoint, neural networks can sometimes capture quantum states with far fewer parameters than required by a full density matrix. For instance, an RBM with few hidden nodes may efficiently represent highly entangled states that otherwise would need exponentially many parameters. Similarly, ansätze based on tensor networks efficiently represent low-entanglement states. Although neural networks do not inherently obey physical entropy laws, with proper regularization or architectural bias they can adapt to the relevant complexity. Empirically, neural network tomography has achieved promising results even for structured but complex states.

### 3.5.1. INTEGRATION IN SPIKING VARIATIONAL AUTOENCODER (SVAE)

In this thesis, SIC-POVMs play a foundational role in constructing the input dataset for SVAE-based quantum state tomography. The spiking encoder processes one-hot encoded SIC outcomes over time to generate a latent representation. The decoder then reconstructs the measurement statistics, which are used to approximate the original density matrix. The symmetry and minimal redundancy of SIC-POVMs reduce the input dimensionality while preserving information, facilitating efficient and accurate inference.

### 3.5.2. SVAE FOR QST

The SVAE is a deep generative model that provides a probabilistic framework for representation learning. An SVAE (often simply called a Variational Autoencoder or VAE) comprises two main components: an *encoder* (or recognition network) and a *decoder* (or generative network). Together, these components model the distribution of data by introducing latent random variables $z$ that capture the underlying factors of variation in the observed data $x$.

In our context, one can view $x$ as representing measured quantum outcomes (or their distribution), and $z$ as a latent representation related to the quantum state. The typical generative process assumes that a latent variable $z$ is drawn from a simple prior $p(z)$ (often a standard multivariate normal distribution) and, conditioned on $z$, the decoder produces a distribution $p_\theta(x|z)$ for the observed data. Here, $\theta$ denotes the decoder parameters (the weights and biases of a neural network). The encoder, parameterized by $\phi$, takes an observation $x$ and returns an approximate posterior $q_\phi(z|x)$ over the latent variable.

Training the VAE involves maximizing the **evidence lower bound (ELBO)** on the data log-likelihood. For a single data point $x$, the ELBO is

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\ln p_\theta(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x)\,\|\,p(z)\right),$$

where $D_{\mathrm{KL}}$ is the Kullback–Leibler divergence. The first term measures the quality of reconstruction (i.e. how well the decoder recovers $x$ from $z$), while the second term regularizes the encoder's distribution so that it remains close to the prior $p(z)$.

The term "stochastic" arises from the sampling of $z$ during training (facilitated by the reparameterization trick, e.g. $z = \mu_\phi(x) + \sigma_\phi(x)\odot\epsilon$ with $\epsilon\sim\mathcal{N}(0,I)$). Both the encoder and decoder are typically deep neural networks.

### 3.5.3. APPLICATION TO QST

In QST, the *data* are the measurement outcomes collected from many copies of an unknown quantum state. Suppose we measure many instances of the state in various settings to obtain a collection $\{x^{(k)}\}$. The goal is to infer the quantum state $\rho$ that generated these outcomes. One approach is to build an SVAE where the decoder $p_\theta(x|z)$ is designed to mimic the quantum measurement process for a given latent representation $z$. In other words, $p_\theta(x|z)$ can be constructed to output probabilities corresponding to $\mathrm{Tr}(E_i\,\rho(z))$ for some parameterization $\rho(z)$ of a density matrix. The encoder $q_\phi(z|x)$ then infers a distribution over possible states $z$ that could have produced the observed outcomes.

Typically, in tomography the assumption is that a single latent $z$ (corresponding to the true state) generates all the measurements; this is akin to a variational Bayes treatment where one updates a prior $p(z)$ based on the aggregated data. For example, for a single qubit one might let $z$ parameterize the polar and azimuthal angles $(\theta,\phi)$ on the Bloch sphere, with the decoder mapping these angles to the outcome probabilities in different measurement bases.

The SVAE framework is advantageous because the KL regularization mitigates overfitting to noise and provides a smooth, low-dimensional latent representation

of $\rho$. Moreover, after training, one can sample from the latent space to obtain an ensemble of likely quantum states, facilitating uncertainty quantification on observables. From an implementation standpoint, training an SVAE involves numerous matrix operations and non-linear function evaluations,tasks that are well suited for memristive neuromorphic hardware. In our envisioned system, both the encoder and decoder networks would be mapped to silicon-based ReRAM arrays, enabling in situ weight updates and real-time processing.



## 3.6. CONCLUSION OF THIS CHAPTER

THE fusion of neuromorphic hardware based on memristor devices with advanced neural network techniques such as spiking variational autoencoders offers a promising route for efficient quantum state tomography. By executing the necessary linear algebra operations directly in memory, our hybrid system minimizes data movement and power consumption while enabling real-time reconstruction of quantum states. This interdisciplinary approach not only advances the field of quantum tomography but also paves the way for broader applications of neuromorphic computing in quantum information science. The theoretical frameworks developed here underpin the experimental fabrication and characterization of memristive devices discussed in Chapter 4.

# 4

# MATERIALS AND DEVICES

*THIS research explores the use of liquid-phase silicon (L-Si) solidified with UV-light and direct wirting with Electron-Beam-Induced Processing (EBIP) to fabricate silicon nanostructures from liquid-phase (heavy polysilanes) in a controllable way, addressing the limitations of traditional high-temperature silicon processing and the large feature size. By utilizing polysilanes as liquid silicon sources, this study demonstrates that EBIP can achieve silicon fabrication at room temperature without requiring annealing. The experimental results show reduced contamination, control over silicon nanostructure, and promising silicon transformation. This novel approach offers significant potential for scalable, cost-effective semiconductor manufacturing and could revolutionize the production of advanced nanodevices for future electronics.*

The contents of this chapter were based on the unpublished paper [74].

## 4.1. INTRODUCTION OF THIS CHAPTER

SILICON continues to play a pivotal role in semiconductor technology due to its exceptional electronic properties, making it essential for a wide array of devices, including transistors, integrated circuits, and memory components [29, 75]. Traditional silicon fabrication methods, such as Chemical Vapor Deposition (CVD) and Molecular Beam Epitaxy (MBE), typically require high processing temperatures and intricate setups, significantly restricting their applicability, substrate compatibility, and overall production efficiency [29, 75].

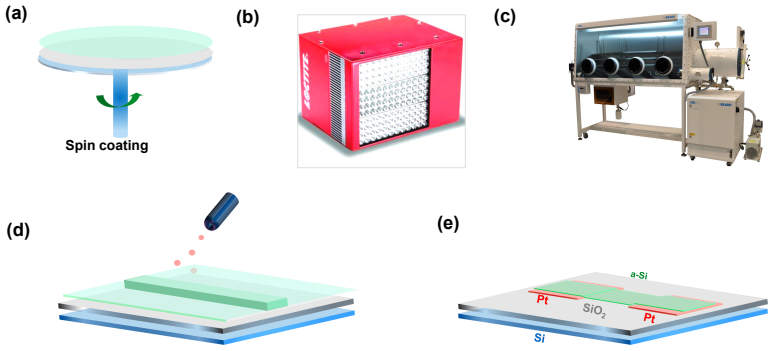
In response to these challenges, liquid-phase silicon (L-Si) precursors, particularly polysilanes, have attracted attention due to their ease of handling and potential for low-temperature fabrication. Polysilanes are silicon-hydrogen compounds that can be spin-coated onto substrates and subsequently transformed into solid silicon films by thermal or laser-induced curing [76–78]. However, the conventional thermal curing process for polysilanes typically necessitates elevated temperatures around 400°C to facilitate dehydrogenation and silicon crystallization, severely limiting their compatibility with temperature-sensitive substrates [76]. Laser-based curing methods, although operating at comparatively lower temperatures, usually suffer from poor spatial resolution, thus limiting their suitability for nanoscale device fabrication [79].

Electron Beam-Induced Processing (EBIP), especially when combined with UV-light pre-treatment, emerges as a highly promising non-thermal alternative capable of fabricating silicon nanostructures directly from liquid polysilane precursors at room temperature. EBIP employs a precisely focused electron beam to selectively cleave Si–H bonds and induce the formation of stable Si–Si networks at nanoscale resolution, effectively overcoming the limitations associated with traditional methods [30]. Notably, EBIP offers exceptional spatial resolution, potentially down to a few nanometers, significantly exceeding the capabilities of conventional thermal or laser-based processes [80]. Furthermore, EBIP substantially reduces impurity incorporation and contamination, addressing another critical drawback of traditional electron beam techniques [81].

The primary objective of this research is thus to assess the feasibility of EBIP as a practical method for the room-temperature direct writing and fabrication of high-quality silicon nanostructures from liquid-phase polysilanes. This investigation targets the fundamental limitations of traditional silicon fabrication techniques by leveraging EBIP's capacity to achieve low-temperature processing, nanoscale precision, and minimal contamination.

Experimentally, this study involves spin-coating polysilane films onto $Si/SiO_2$ substrates and platinum electrodes, followed by exposure to electron beams under systematically controlled parameters including electron beam voltage, current, and exposure duration. The resulting silicon nanostructures are analyzed for their morphological characteristics, electrical performance, and stability to elucidate the relationships between processing conditions and material properties.

By successfully demonstrating EBIP-based fabrication at room temperature, this research could significantly broaden the range of compatible substrates and enable cost-effective, high-resolution silicon nanostructure production, potentially




Figuur 4.1: **(a)**. Schematic of spin-coating of L-Si. **(b)**. Photo of UV source (Loc-tite LED Flood UV). **c**. Glovebox for a moisture-/air-free environment for L-Si film preparation. **d**. Schematic illustration of Electron Beam-Induced Processing (EBIP) on L-Si films. **e** fabricated devices stacks and configuration.



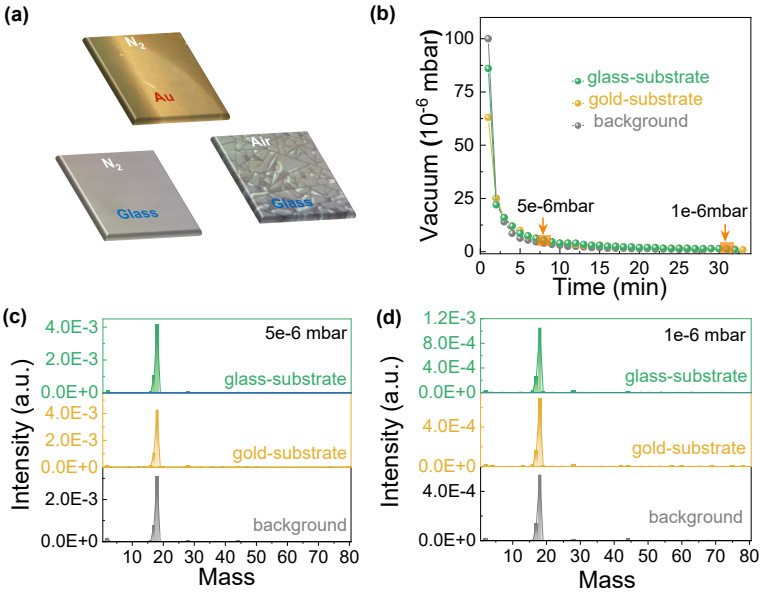
revolutionizing applications in advanced semiconductor devices, sensors, and memory technologies.

## 4.2. METHODOLOGIES

As shown in Fig. 4.1, it demonstrates the fabrication process, including schematics and images that describe the process of fabricating and configuring silicon-based nanostructures. 25v% heavy polysilanes (chemical compounds of heavy polysilanes have been listed in the table 4.2) diluted by cyclooctane solution and spin-coated on various substrates, including glass-/Au-/Si-based wafers, at 3000 rpm. It is followed by UV light illumination using the Loctite LED Flood UV (Fig. 4.1b) with a power of 300 mW/cm$^2$ for 15 minutes to cure and solidify liquid-phase polysilanes.

To ensure that the spin-coated thin films can be operated in the high-vacuum chamber of a focused electron beam or scanning electron microscope (SEM). Residual Gas Analysis (RGA) for monitoring gas release during the process was processed using a residual gas analyzer rugged mass spectrometer (Inficon Transpector) connected with a Turbo pump for 30 minutes.

A focused ion beam scanning electron microscope (FIB-SEM, ThermoFisher Hydra Bio Plasma-FIB) is adopted for writing the Si features on L-Si and morphological/elemental analysis (EDS). beforehands, UV-cured L-Si thin film was transfered into FIB-SEM with protection of $N_2$ flushing. The schematic in Fig. 4.1(d) illustrates the use of an electron beam (E-beam) to write silicon nanowires on a L-Si film. This process involves precise patterning at the nanoscale, crucial for creating nanoscale devices or structures. Scanning Transmission Electron Microscopy equipped with elemental energy loss spectroscopy (STEM-EELS) was applied to analyze the components of Si features after being written with electron beams. STEM lamellae were prepared with a focused ion beam using an FEI Helios Dual Beam system and was thinned to electron transparency. The samples were imaged




Figuur 4.2: a) Images of spin-coated liquid polysilanes on different substrates (glasses and gold) and in different atmospheres ($N_2$ and air). b) The pressure change under vacuum with time for different strbusted polysilanes, c) Mass analysis under 5e-6 mbar on glass-/gold-substrate polysilanes, d) Mass analysis under 1e-6 mbar on glass-/gold-substrate polysilanes

with a ThermoFisher Titan system with an accelerating voltage of 200 kV. The STEM imaging modes used are annular dark-field STEM (STEM-ADF) and annular bright-field STEM (STEM-ABF).

The EBIP/L-Si route operates under distinct boundary conditions—*room temperature*, inert ambient, localized energy delivery, and post-CMOS compatibility—which shift the optimization targets away from furnace throughput toward *energy per feature* and *thermal-budget compliance*. A direct metric-for-metric benchmark with high-temperature CVD/MBE or plasma methods would therefore be premature. Instead, we position EBIP/L-Si against representative high-$T$ and low-$T$ options along decision criteria relevant to integration as illustrated in the table 4.1.

Electrical measurements for assessing the resistivity of the silicon nanostructures were conducted using B1500A euiqped with a probe station.

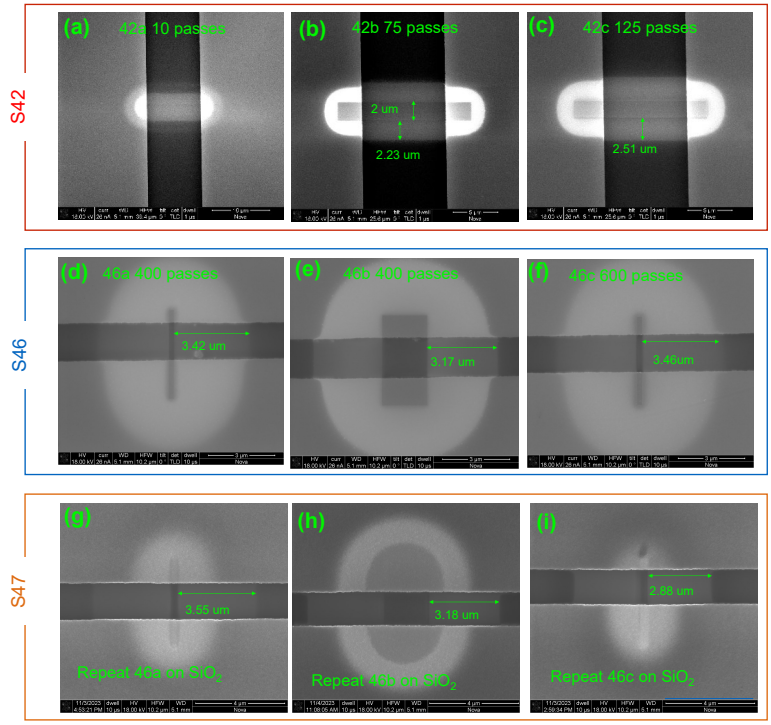
## 4.3. RESULTS AND DISCUSSION

### 4.3.1. RGA ANALYSIS

THE set of figures presents results from an experiment comparing the residual gas analysis (RGA) of different substrates (glass and Au) under pre-processing environmental conditions. This RGA information is crucial when understanding the

Tabel 4.1: Criteria-based positioning of EBIP/L-Si vs. representative methods (scope statement) [26, 82–90].

| **Criterion** | **EBIP/L-Si** | **High-$T$ CVD/MBE** | **Low-$T$ plasma/ALD** |
|---|---|---|---|
| Thermal budget (max °C) | RT / BEOL-safe | 400–1000 °C | 100–300 °C |
| Ambient/precursors | Inert, minimal residues | Reactive gases; higher contamination risk | Plasma/precursors; byproducts possible |
| Substrate compatibility | Flexible/BEOL stacks | Limited by high $T$ | Wider than high-$T$ |
| Energy delivery | Localized (e-beam) | Global (furnace) | Global/semi-local |
| Feature size/placement | Sub-100 nm, direct write | Defined by lithography + deposition | Lithography + conformal films |
| Throughput (today) | Low–moderate (direct write) | High (wafer-scale) | Moderate–high |
| Integration overhead | Low thermal stress; post-CMOS friendly | High thermal stress; anneals | Moderate; plasma exposure |
| Typical failure modes | e-beam residue, local defects | Dopant/strain diffusion, film stress | Plasma damage, impurities |

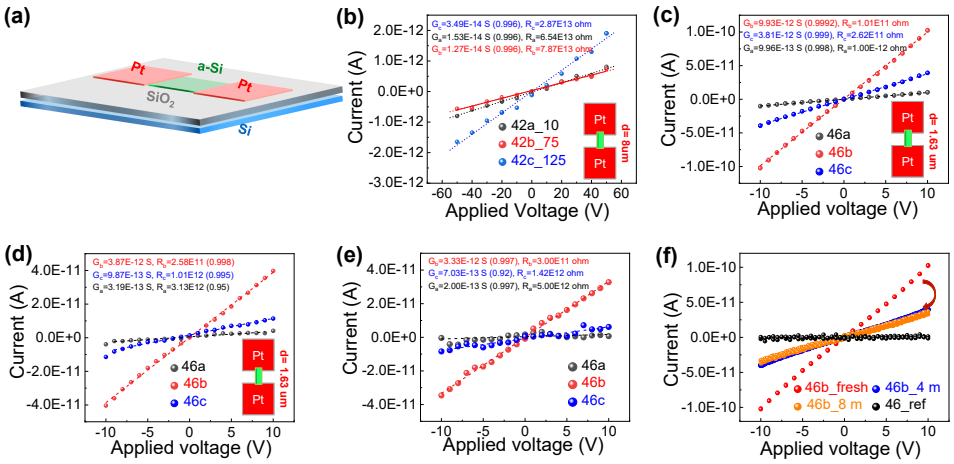



Figuur 4.3: an overview of samples with a dose of 0.009 pC (**a**) and a dose of 0.07 pC (**b**) and a dose of 0.1 pC (**c**) for sample 1 (denoted as "S1a", "S1bänd "S1c", respectively). an overview of samples with a dose of 13.0 pC for 0.2 $\mu$m × 4.0 $\mu$m pattern (**d**) and a dose of 13.0 pC for the 2.0 $\mu$m × 4.0 $\mu$m pattern (**e**) and a dose of 15.6 pC for the 0.2 $\mu$m × 4.0 $\mu$m pattern (**f**) for sample S2 (denoted as "S2a", "S2bänd "S2c", respectively). For S3 (g, h, i), they are repeating EBIP experimentation from S2a, S2b and S2c without polysilanes on the top of substrate.

stability and suitability of these materials for processes like electron-beam-induced processing (EBIP) or other vacuum-based fabrication techniques in case the pre-processed film will trigger unsafe issues and contaminations.

From Fig. 4.2a, it can be seen that the L-Si surface conditions were spin-coated on different substrates and in different atmospheres. It displays one Au substrate (top) and two glass substrates (bottom two) thin films in ($N_2$) environment (ones at the top and bottom left) and in an air environment (the one at the bottom right). In the bottom-right image, the same L-Si film on a glass substrate was observed after being exposed to air for 10 minutes.

From the above observations, the L-Si thin films on a glass substrate exposed to air show visible surface changes, which could indicate oxidation. This contrasts with the nitrogen environment, where the surface looks clean and unaffected as presented in Fig. 4.2(a). Both Au-/glass-substrate thin films remain shiny and clean in a nitrogen environment, suggesting that Au and glass have good adhesion features




Figuur 4.4: (a) Schematic cross-section of the Pt/a-Si/$SiO_2$/Si device structure, showing platinum electrodes on top of amorphous silicon layer. (b)–(e) Representative current–voltage (I–V) characteristics for various samples (e.g., S1, S2), measured from −60 V and −10 V to +10 V. The data are plotted on a semi-log scale to highlight the extremely low currents and reveal the linear (ohmic) conduction regimes. Dashed lines in each plot indicate linear fits, from which conductances on the order of $10^{-14}$ S (and resistances $\sim 10^{13}\,\Omega$) are extracted. (f) Comparison of I–V responses for stability over 8 months. All data sets exhibit nearly symmetric behavior around 0 V, consistent with ohmic conduction dominated by the bulk properties of the a-Si layer. The insets in (b)–(e) schematically show the measurement geometry with Pt electrodes contacting the a-Si film.

for L-Si, which can both perform further device fabrication on them. Figure 4.2b is a pressure *vs.* time plot. It tracks the pressure change over time (in minutes) for both the glass and Au substrates compared to a background reading in the vacuum chamber of the residual gas analyzer. The experiment was conducted at two different vacuum levels: 5e-6 mbar and 1e-6 mbar. For the glass/Au-substrate L-Si thin films, they exhibit a lower initial pressure that rapidly decreases as time progresses, indicating negligible outgassing during the first few minutes, which then stabilizes. For background curves (black line), it serves as a control and remains consistently similar to both aforementioned substrates, indicating that the pressure changes are largely the same as the substrates without any L-Si and therefore no or negligible release of the gaseous or liquid phase of heavy polysilane molecules. After zooming in on the plots with pressure of 1e-6 mbar, we can still find some trace molecules, and they are listed in Table 4.4, which is mostly due to the solution residuals and negligible breaking links of polysilanes from the L-Si films.

The Fig. 4.2 (c,d) demonstrates that there is no noticeable release of gas molecules under both vacuum levels, resulting in higher pressure readings. There are also no massive trapped gases in the glass/Au substrate and reactions with its surface. Meanwhile, the concern of contamination from outgassing in an electron beam

vacuum environment is eliminated from this experiment.

### 4.3.2. EBIP WRITING FOR THE L-SI TRANSFORMATION

THIS study investigates the influence of electron beam irradiation on L-Si films. Scanning electron microscopy (SEM) was employed to visualize the morphological changes induced by electron-beam-induced processing (EBIP). The experiments are organized into three series: S1, S2, and S3, each representing a different set of irradiation conditions with varying numbers of electron beam passes (as shown in Fig. 4.3).



| **Species** | **$T_B$ (K)** | **Concentration (MSDS)** |
|---|---|---|
| $SiH_4$ | 161.3 | 0–5% |
| $Si_2H_6$ | 259 | 0–5% |
| $Si_3H_8$ | 326 | 0–5% |
| $Si_4H_{10}$ | 381 | 10–50% |
| $Si_5H_{12}$ | 467/426.2 | 10–50% |
| $Si_6H_{14}$ | 500/467 | 10–50% |
| $Si_7H_{16}$ | 600/470 | 10–50% |
| $Si_xH_y$ (x≥ 8) | 700 | N/A |

Tabel 4.2: Chemical components of heavy polysilanes from MSDS sheet

For S1, as shown in Fig. 4.3(a-c), it shows the result of 0.09-0.1 pC of electron beam exposure on an L-Si-coated substrate. The electron beam-irradiated region is narrow, and there is a bright contrast surrounding the central region, possibly resulting from the presence of a higher intensity or density of material.

The variation in brightness between the central and peripheral areas of the irradiated region can be attributed to several factors [91]: i) The darker appearance is likely due to electron beam-induced charging, sample damage, or surface modifications including etching effects from the electron beam on L-Si films; ii) The enhanced brightness at the edges is possibly due to edge effects. Key contributing factors include: a. The angle of incidence at the boundaries enhances the escape of secondary electrons, leading to increased brightness; b. Scattered electrons at the edges interact more with the sample or escape more readily from thinner layers, thereby increasing the secondary electron yield; c. Electrons incident through L-Si films scatter back to the L-Si films after reflecting from the $Si/SiO_2$ substrate, where they incur the L-Si phase change, and thus it makes the L-Si more conductive. c. Some electron beam energy may spill over into adjacent regions, further modifying the surface and enhancing the brightness at the periphery.

In Fig. 4.3d-e, they demonstrate the result of 400 electron beam passes (a dose of 13.0 pC), with a peripheral width of 3.42 $\mu$m and 3.17 $\mu$m, respectively. The

structure is well defined, with a clear boundary between the exposed and unexposed regions. The deposited material appears to have a relatively uniform width. For 4.3f, it shows the result of 600 electron beam passes (a dose of 15.6 pC), where the edge structure is 3.46 $\mu$m wide. The increased number of passes results in a more intense film growth, with the structure becoming thicker and more pronounced.

The S2 series shows the effect of a larger number of electron beam passes (400 and 600) on the silicon nanostructures processing. The figures show that increasing the number of passes leads to a slight increase in the width of the edge area. However, the process seems to plateau, as the difference between 400 and 600 passes is relatively small. This suggests that after a certain number of passes, additional beam exposure may result in diminishing returns in terms of peripheral area width. Furthermore, it improves the density and thickness of the material, making it useful for applications where uniformity and consistency are required. For experiments S3 without polysilanes (g, h, i) on $SiO_2$, it shows the results of exposure to the electron beams on a substrate without L-Si films. The images illustrate no significant Si growth or modification, confirming that the electron beam alone does not result in material growth without the presence of liquid silicon.



The S3 series experiment validates that the nanostructures formed in previous experiments are indeed a result of the electron-beam interaction with the liquid silicon precursor. The SEM images from S1 and S2 clearly demonstrate that increasing the number of electron beam passes results in progressively wider and more defined nanostructures. This provides a high level of control over the dimensions of the deposited material, making EBIP a powerful tool for nanoscale fabrication.

The ability to control the size, width, and density of the nanostructures by adjusting the number of electron beam passes makes EBIP with L-Si a promising technique for applications in semiconductor devices and nanotechnology. This approach could be particularly useful for fabricating components like transistors, sensors, and memory devices, where precision and material purity are crucial. The combination of these images illustrates the power of EBIP in producing high-resolution, controllable silicon nanostructures and confirms the role of liquid-phase silicon as a precursor in the process.

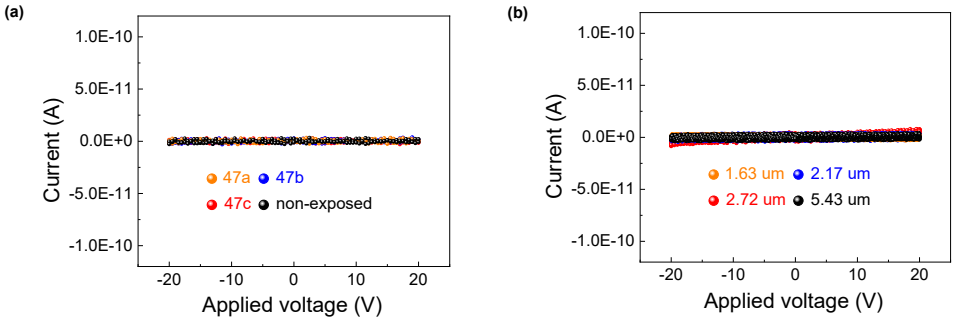
Figure 4.4 presents the experimental study of electrical transport in devices composed of Pt/a-Si/Pt (platinum and amorphous silicon) structures. The data is analyzed to determine the conductance (G), resistance (R), and the effects of device geometry and aging.

The device consists of substrates including an $SiO_2$ layer on a silicon (Si) base and two platinum (Pt) contacts positioned on the top of the amorphous silicon (a-Si) layer (as illustrated in Fig. 4.4a). The inter-electrode distance (labeled as "d") is an essential geometric parameter impacting resistance. The graph shows three sets of data points (red, blue, and black), representing different experimental conditions. From linear fitting, we analyzed the conductance ($G$) and resistance ($R$).

Current-voltage (I-V) characteristics of three samples(*1a_10*, *1b_75*, and *1c_125*), as shown in Fig. 4.4b, are measured between $-60$V and 60V. Each set of data points (gray, red, and blue) corresponds to a different sample (eg., S1a,

Tabel 4.3: Summary of patterns with their respective dimensions, passes, accelerating conditions, and calculated electron beam doses.

| Patterns | Size (um) | Passes | Accelerating Voltage & Current | PitchXY (nm) | HFW (um) | Resistivity (nΩ.m) | Dose (pC/nm$^2$) |
|---|---|---|---|---|---|---|---|
| 46a | 4x0.2 | 400 | 18kV,26nA | 4x4 | 10.2 | 6.27E12 | 13.0 |
| 46b | 4x2 | 400 | 18kV,26nA | 4x4 | 10.2 | 6.34E12 | 1.3 |
| 46c | 4x0.25 | 600 | 18kV,26nA | 4x4 | 10.2 | 2.06E12 | 15.6 |
| 42a | 15x2 | 10 | 18kV,26nA | 4x4 | 39.4 | 8.58E14 | 0.0087 |
| 42b | 15x2 | 75 | 18kV,26nA | 4x4 | 25.6 | 8.27E14 | 0.0650 |
| 42c | 15x2 | 125 | 18kV,26nA | 4x4 | 32 | 3.53E14 | 0.1083 |
| ref. | 50x50 | 0 | 0kV,0nA | 0 | N.A. | 8.50E16 | 0.0 |




Figuur 4.5: **a** Control experimentation for no L-Si film on substrates. **b** Control experimentation for L-Si film on substrates without electron-beam irradiation.

S1b,S1c, respectively), while the dashed lines show linear fitting used to extract the conductance $G$ and resistance $R$ values. The extracted conductances are on the order of $10^{-14}$ S (e.g., $G_c = 3.49 \times 10^{-14}$ S), yielding resistances on the order of $10^{13}\,\Omega$. The correlation coefficients are all approximately 0.996, indicating an excellent fit to a simple ohmic (linear) model. The small inset schematic illustrates the device geometry, with platinum (Pt) electrodes and an active layer (L-Si induced Si) of thickness $d \approx 8\,\mu$m. The nearly symmetric I-V response around 0V further supports ohmic conduction dominated by the bulk properties of the material. From Fig. 4.4b, it can be concluded that the conductance increases with increasing number of electron beam passes.

In Fig. 4.4c, it demonstrates that the device S2b exhibits higher conductance ($G_b = 9.93 \times 10^{-12}$ S) compared to S2c ($G_c = 3.81 \times 10^{-12}$ S).The differences in conductance indicate sensitivity to device geometry (exposed pattern dimensions). The conductance decreases as the electrode spacing increases. This result aligns with classical resistive models where $R = \rho \cdot (d/A)$ ($\rho$: resistivity, $A$: cross-sectional area). Data is labeled for different device conditions: *fresh* (Fig. 4.4c), *4 months* (4m: (Fig. 4.4d)), and *8 months* (8m: (Fig. 4.4e)) for S2, and the comparison of S2b aging results was plotted in Fig. 4.4d, where it presents resistance increases over time (fresh: lowest resistance; 8m: highest resistance) and stabilized. Aging is likely induced by oxidation, structural defects, or changes in the amorphous silicon layer,

leading to increased resistivity. The temporal changes highlight the device's stability under environmental conditions. $G_s$ and $R_s$ values are consistent with expected trends for amorphous silicon, where electron conduction is limited by defects and trap states. Lower $d$ or fresher devices yield higher conductance. High correlation coefficients ($R^2 > 0.99$) in all linear fits suggest that transport is dominated by ohmic (linear) conduction, with minimal non-linear effects.

As a reference, S3, along with non-exposed substrate, is showing extremely high resistance, as illustrated in Fig. 4.5a. Meanwhile, I-V curves for different spacings of d for L-Si-coated substrates are also studied, as shown in Fig. 4.5b, which presents similar results to Fig. 4.5a. It implies that electron beam irradiation is the key factor to make some difference on the resistivity of the L-Si film.



Tabel 4.4: Potential molecules associated with various mass ranges released under high vacuum level.

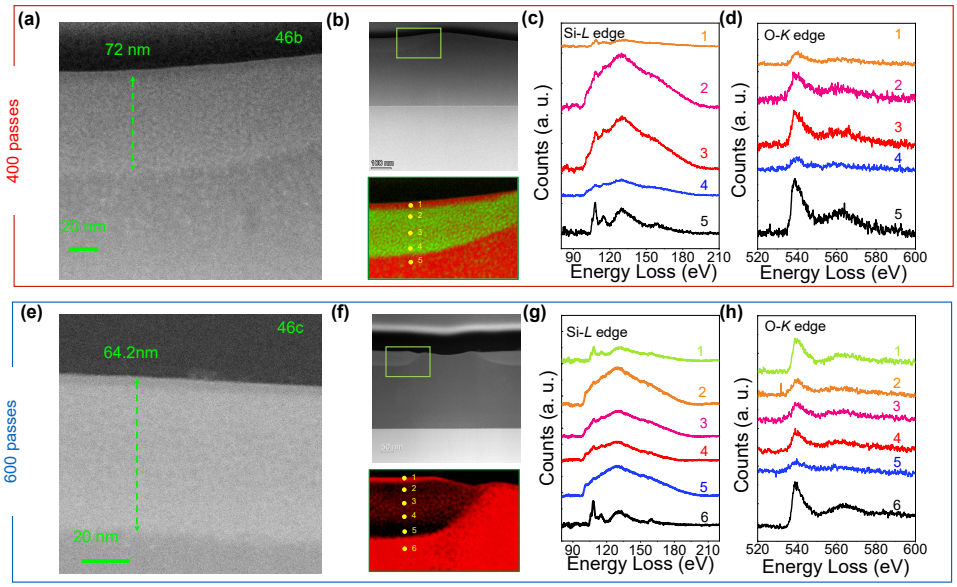
| **Mass** | **51,52** | **57,58** | **66,67,69,72** | **75,78,79,81** | **87,90,95,97,98** | **101,102,110,113,116,117** |
|---|---|---|---|---|---|---|
| **Possible chemicals** | $C_4H_3$, $C_4H_4$ | $Si_2H$, $Si_2H_2$ | $C_5$-$H_6$/$H_7$/$H_9$/$H_12$ | $C_6$-$H_3$/$H_6$/$H_7$/$H_9$ | $Si_3$-$H_3$/$H_4$/$H_5$/$H_6$ | broken cyclooctane |

Increasing electrode spacing ($d$) increases resistance, which is consistent with theoretical resistive scaling. Over time, environmental factors degrade the material, leading to increased resistance and decreased performance. The amorphous silicon layer shows trap-limited conduction with sensitivity to defects and fabrication quality. Aging effects and the sensitivity of electrical properties to environmental conditions indicate potential challenges for long-term device stability.

To understand the difference of conductance of different samples. *HRSTEM-EELS* was applied as illustrated in Figure 4.6.It compares the microstructural and compositional properties of two samples, 2b (400 passes) and 2c (600 passes). In Fig. 4.6(a) and Fig. 4.6(e), the low-magnification *HRSTEM* cross-sections reveal that the film thickness decreases from about 72 nm to 64 nm when the number of passes is increased from 400 to 600. This reduction may be attributed to a more efficient densification or partial etching during the additional passes. High-resolution TEM images Fig. 4.6(b) and Fig. 4.6(f) show smooth interfaces and uniform contrast, indicative of homogeneous film deposition with no visible interfacial voids or large-scale phase separation.

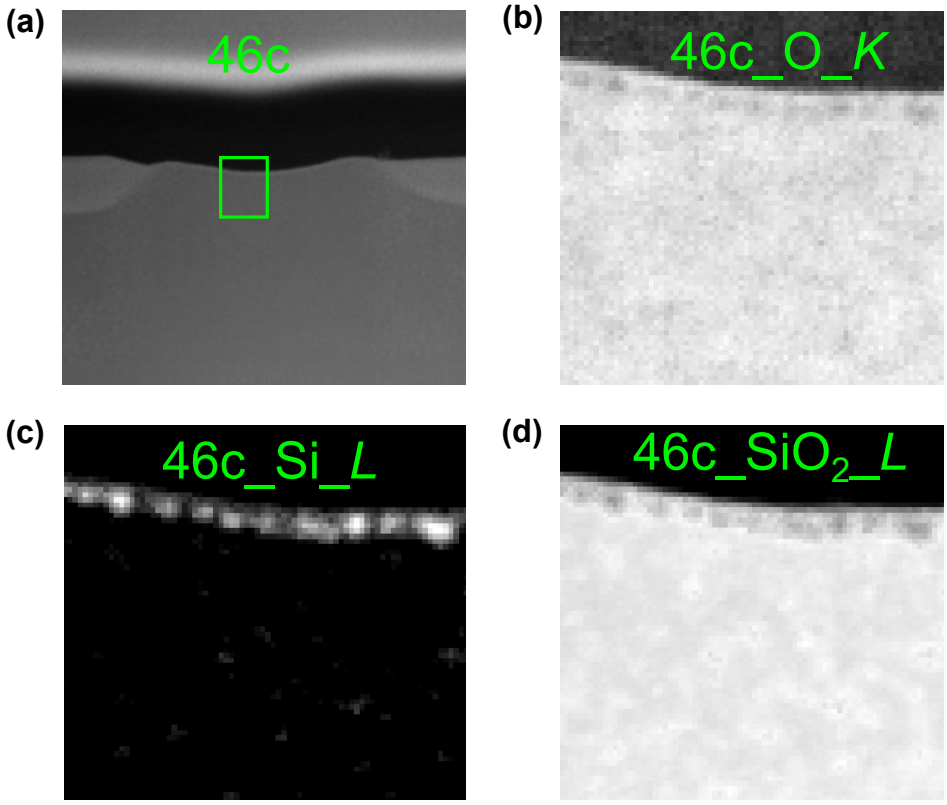
The EELS spectra in Fig. 4.6(c-d) and Fig. 4.6(g-h) were collected at multiple locations across each film (labeled 1–4), focusing on the Si-$L$ and O-$K$ edges. In both samples, the Si-$L$ edge exhibits characteristic peaks around 100–120 eV, consistent with silicon in an amorphous or partially oxidized state, while the O-$K$ edge near 530 eV indicates the presence of oxygen throughout the film. The spectral shapes are nearly identical across each film thickness, suggesting that the local chemical environment remains uniform laterally and through the depth of the film. Overall, the comparison between S2b and S2c indicates that increasing the number of passes modestly reduces the film thickness without significantly altering the Si-O bonding or the compositional uniformity observed in these *HRSTEM-EELS* measurements.

Figure 4.7 provides a closer look at the local compositional uniformity in sample




Figuur 4.6: **(a)–(d)** Cross-sectional high-resolution *STEM* and electron energy-loss spectroscopy (EELS) analysis of sample S2b. **(a)** Cross-section overview of S2b and the interested area in the green dashed line framed under the high-resolution *STEM*. **(b)** High-resolution STEM image of the region of interest, showing a uniform film thickness of approximately 72 nm (top), with an overlaid color *HRSTEM-EELS* map (bottom) indicating elemental distribution derived from *HRSTEM-EELS*. The dotted region highlights the area where the spectral maps in **(c)** and **(d)** were acquired. **(c)** EELS spectra at the Si-*L* edge from five different locations (curves 1–5). **(d)** Corresponding EELS spectra at the O-*K* edge. **(e)–(h)** Cross-sectional high-resolution STEM/EELS results for sample S2c. **(e)** Cross-section overview of S2c and the interested area in the green dashed line framed under the *STEM*. **(b)** High-resolution STEM image of the region of interest, showing a uniform film thickness of approximately 64 nm (top), with an overlaid color *HRSTEM-EELS* map (bottom) indicating elemental distribution derived from *HRSTEM-EELS*. The dotted region highlights the area where the spectral maps in **(g)** and **(h)** were acquired.

2c (600 passes), complementing the cross-sectional *HRSTEM/EELS* data shown in Fig. 4.6(e). In Fig. 4.7(a), the low-magnification cross-sectional image reveals a smooth interface between the film and substrate. The green box highlights the region where the elemental maps in Fig. 4.7(b, c, d). Fig. 4.7(b) shows the O-*K* edge map, demonstrating that oxygen is present throughout the film thickness with minimal lateral variation except for the topmost layer. In Fig. 4.7(c) (*HRSTEM-EELS*), the Si-*L* edge map confirms that silicon is uniformly distributed on the topmost layer. The additional map in Fig. 4.7(d) presents *HRSTEM-EELS* $SiO_2$-*L* mapping, emphasizing more fully oxidized silicon states (i.e., $SiO_2$), but the topmost layer shows the presence of $SiO_x$.




Figuur 4.7: **(a)** Cross-sectional *HRSTEM-EELS* image of sample 2c (600 passes), showing the thin film atop the substrate and a highlighted region of interest (green box). **(b)** Elemental map at the O-*K* edge, revealing the spatial distribution of oxygen within the film. **(c)** Corresponding map at the Si-*L* edge, illustrating silicon-rich regions. **(d)** Map of *2c_SiO$_2$_L* highlighting regions of more oxidized silicon (i.e., $SiO_2$).

Therefore, these maps confirm that the topmost layer of exposed thin film is $Si/SiO_x$, which is incured by electon beams. Combined with the *HRSTEM-EELS* spectra in Fig. 4.6, these results indicate that increasing the number of passes (from 400 to 600) modestly alters the film thickness but does not significantly change the Si-O bonding environment or uniformity, where the modified area is oxidized naturally after being exposed to air. Such compositional stability suggests that the processing can reliably produce uniform amorphous films for device applications. In both samples (2b and 2c), the EELS data reveal similar elemental composition profiles, suggesting that increasing the number of passes from 400 to 600 slightly reduces the film thickness (72 nm to 64 nm) but does not significantly alter the Si and O bonding environments, which is probably resulting from the etching effect of a higher dose of electron beam. The resistivity of electron beam-cured L-Si films

Tabel 4.5: Summary of pattern dimensions, passes, and operating conditions ($V_{Acc}$ & $I_{Current}$).

| Patterns | Size (um) | Passes | $V_{Acc}$ & $I_{Current}$ |
|---|---|---|---|
| 42a | 15x2 | 10 | 18kV,26nA |
| 42b | 15x2 | 75 | 18kV,26nA |
| 42c | 15x2 | 125 | 18kV,26nA |
| 46a | 4x0.2 | 400 | 18kV,26nA |
| 46b | 4x2 | 400 | 18kV,26nA |
| 46c | 4x0.25 | 600 | 18kV,26nA |
| 47a | 4x0.2 | 400 | 18kV,26nA |
| 47b | 4x2 | 400 | 18kV,26nA |
| 47c | 4x0.25 | 600 | 18kV,26nA |
| ref. | 50x50 | 0 | 0kV,0nA |

Tabel 4.6: Electrical and processing parameters for patterns 2b, 2c, and resistivity change over time once exposed to air. The area calculation is based on the measured thickness and dimensions of the patterns.

| Patterns | Area (nm·$\mu$m) | Passes | $\rho$(fresh) (n$\Omega$·m) |
|---|---|---|---|
| S2b | 450.72 | 400 | 2.84e13 |
| S2c | 444.26 | 600 | 7.29e13 |
| Ref. ($SiO_2$)[a] | N.A. | N.A. | 1.00e23 |
| Ref. (Si)[b] | N.A. | N.A. | 2.30e12 |

[a] Reference for resistivity of $SiO_2$ [92].
[b] Reference for resistivity of Si [93].

will decrease to a great extent as the number of passes/doses increases, as shown in table 4.3, which illustrates that the polysilanes can be used as printing ink for Si nanostructure directly writing with the EBIP technique. This is also confirmed in Table 4.6, which shows that the stability test of the electron-beam-induced Si or SiO**x** shows long material stability after being exposed to the air for a while. It also indicates that the resistivity of the Si or SiO**x** is quite close to that of intrinsic Si, which provides a valid pathway to a high-precision, cost-effective technique for future manufacture of flexible electronics and thin-film transistors.

## 4.4. CONCLUSION OF THIS CHAPTER

IN conclusion, this research presents a novel approach to overcoming the challenges of liquid-/gasous-phase silicon precursor processing by leveraging the capabilities of EBIP to convert liquid-phase silicon into high-resolution nanostructures at room temperature. By eliminating the need for liquid-/gasous-phase processing and

offering a low-contamination, highly precise fabrication method, EBIP represents a significant advancement in the field of silicon semiconductor technology. The findings of this research have the potential to impact the future development of nanoscale devices, contributing to the evolution of semiconductor manufacturing in an increasingly demanding technological landscape.

# 5

# FORMING-FREE OXRERAM DEVICES

*MEMRISTOR technology shows great promise for energy-efficient computing, yet it grapples with challenges like resistance drift, and inherent variability. For filamentary Resistive RAM (ReRAM), one of the most investigated types of memristive devices, the expensive electroforming step required to create conductive pathways, results in increased power/area overheads and reduced endurance. In this study, we present novel $HfO_2$-based forming-free ReRAM devices, PdNeuRAM, that operate at low voltages, support multi-bit functionality, and display reduced variability. Through a deep understanding and comprehensive material characterization, we unveil the key process that allows this unique behaviour: a Pd–O–Hf configuration that capitalizes on Pd innate affinity for integratinging into $HfO_{2-x}$. This structure actively facilitates charge redistribution at room temperature, effectively eliminating the need for electroforming. Moreover, the fabricated ReRAM device provides tunable resistance states for dense memory and reduces programming and reading energy by 43% and 73%, respectively, for image classification and gesture recognition tasks. This study reveals novel mechanistic insights and delineates a strategic roadmap for the realization of power-efficient and cost-effective ReRAM devices.*

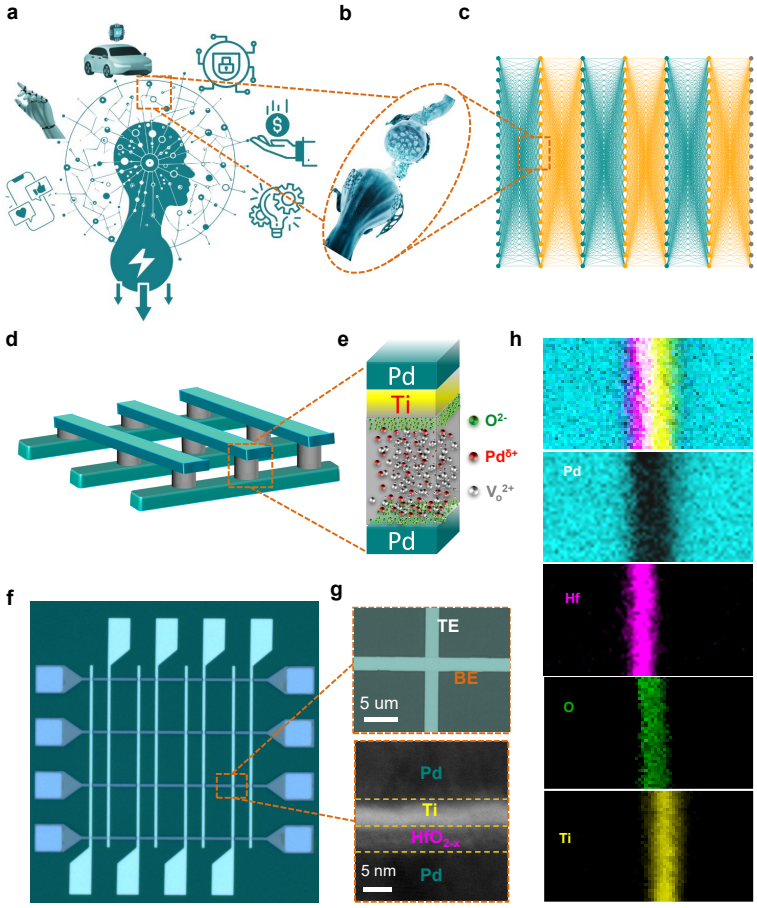
The contents of this chapter were based on the conference paper [94] preprint paper [95], granted papent[96] and in-revision paper [97]

## 5.1. INTRODUCTION OF THIS CHAPTER

IN this emerging era defined by Artificial General Intelligence (*AGI*) and Internet of Things (*IoT*) technologies [32], Computing-in-Memory (*CIM*), where computation and storage seamlessly converge in a single physical locale, has ascended as a visionary next-generation computing paradigm [98]. Among the many candidate technologies for computing cells, Resistive Random Access Memory (*ReRAM*) has captivated both academia and industry with its compelling attributes, including non-volatility, zero leakage power, a simplified structure, high integration density, rapid switching speeds, and inherent *CMOS* compatibility [32, 34, 35]. However, beneath these luminous advantages lie formidable challenges, most notably, the prerequisite for electroforming pre-processing, where the electroforming phase necessitates the application of substantially elevated voltages to establish Conductive Filaments (*CF*) within the resistive medium [35]. It not only intensifies energy consumption and complicates fabrication but also compromises device reliability and endurance [33, 99, 100]. Moreover, the pronounced current density during *ReRAM* write operations further exacerbates concerns about power consumption and long-term durability, often compelling designers to enlarge *ReRAM* cell dimensions, a trade-off that diminishes circuit packing density and inflates manufacturing costs [99, 101].

To address these challenges, $HfO_2$ has been widely explored as a premier dielectric candidate, celebrated for its superior breakdown voltage, robust thermal stability, and exceptional *CMOS* compatibility. These remarkable attributes promise not only a reduction in operational power consumption but also enhanced thermal performance over alternative dielectrics [102–105]. Furthermore, the incorporation of Multi-Level Resistance (*MLR*) technology in $HfO_2$ boosts memory density and slashes manufacturing costs [106–108], offering significant advantages for neuromorphic computing applications. Yet, current specialized fabrication methods, such as thermal annealing [37], X-ray irradiation [35], exotic element doping [37, 38], and plasma treatments [38, 40], are designed to introduce defect states and reduce forming voltages. Aforementioned techniques will increase manufacturing costs and complexity, typically demand substantial energy, and may inadvertently impair transistor performance [109]. Therefore, there is a compelling need for simpler, energy-efficient fabrication processes for forming-free *ReRAM* devices.

In this study, inspired by the remarkable efficiency of biological neural cells, as illustrated in Fig. 5.1a, b, which underpin the acceleration of neuromorphic devices for fully connected artificial neural networks (Fig. 5.1c) [32], we proposed a streamlined, forming-free *ReRAM* fabrication approach that leverages conventional Integrated Circuit (*IC*) manufacturing processes. Notably, the presence of Pd, serving as both an electron injection source and catalytic center in $HfO_{2-x}$, fosters a particular atomic configuration that diminishes charge diffusion barriers [110]. Synergistically, the cooperative action of these integrated stacks culminates in forming-free behavior. Electrical characterization reveals significantly lower initial resistivity in pristine PdHT devices compared to conventional Pt/$HfO_{2-x}$/Ti/Pt (PtHT) devices [111]. Moreover, the PdHT devices exhibit stable multilevel resistance states and low-power switching capabilities, achieved through modulation of *RESET*




Figuur 5.1: **a** An overview of potential applications for neuromorphic systems. **b** Schematics of a biological synapse. **c** Schematics of fully connected artificial neural networks. **d** Schematic illustration of crossbars. **e** schematics of OxReRAM cell. **f** *EELS* mapping of the cross-section of the OxReRAM cell. **g** *SEM* image of as-fabricated 4 × 8 crossbars. **h** *SEM* image of one cell (top) and *STEM* image of cross section of one OxReRAM cell (bottom).

stopping voltages ($V_{RESET,\ stop}$), where the RESET is the operation of switching off the memory devices; the bias at the switching point is defined as V*RESET*; and the sweeping stop voltage is defined as $V_{RESET,\ stop}$. These attributes render the proposed PdHT device a compelling candidate for energy-efficient neuromorphic computing applications. Experimental validations further confirm substantial energy savings during both training and inference phases of Spiking Neural Networks (*SNNs*), demonstrating effective performance in tasks such as image classification and gesture recognition.

Tabel 5.1: The-state-of-the-art forming-free devices

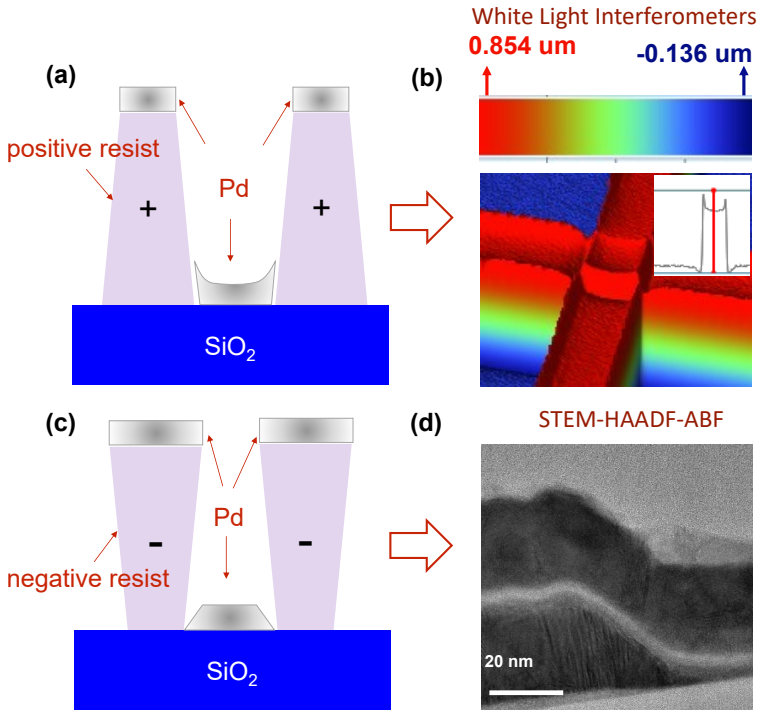
| Types | TE | $Th_{OEL}$[1] | BE | On/Off | $V_{Set,\ Reset}$ | ML[2] | $T_{ann.}$[3] | Treatment |
|---|---|---|---|---|---|---|---|---|
| **no $HfO_2$** | W [112] | 20 nm $WO_x$ | Pd | ~ 20 | 1.0, -3.0 | No | No | oxygen plasma |
| | Ta [39] | 7 nm $Ta_2O_5$ | Pt | ≥ 10 | 1.2, -0.9 | No | 600°C | annealing |
| | $SrTiO_3$ [36] | 400 nm $MoO_3$ | Pt | ~ 10 | 3.5, -3.8 | No | 100°C | proton injection |
| | Al [113] | 150 nm $SiN_x$ | $p^{++}$ Si | ~ 10 | 3.2, -1.5 | No | No | hydrogen plasma |
| | TiN [114] | 5 nm $HfTiO_x$ | TiN | ≥ 30 | 1.0, -1.0 | 2 | 400°C | annealing |
| **$HfO_2$** | Ge [37] | 5 nm $HfO_2$ | Pd | ≥ 500 | 3.2, -0.8 | No | 600°C | Ge doping |
| | Al [115] | 60 nm $HfO_x$ | $p^{++}$ Si | ~ 3 | 2.9, -2.5 | No | 550°C | elements doping |
| | Pt [35] | 20 nm $HfO_2$ | TiN | ≥ 10 | 0.70, -0.70 | No | No | X-Ray |
| | TiN [100] | 3 nm $HfO_x$ | TiN | ~ 10 | 0.75, -0.6 | 2 | 450°C | annealing |
| | Pd | 5 nm $HfO_{2-x}$ | Pd | ~ 19 | 0.56, -0.58 | ≥ 2 | No | No [This work] |

[1] Thickness of oxygen exchange layer.
[2] Multi-level resistance and number of bits.
[3] Temperature of thermal annealing

## 5.2. METHODOLOGIES

### *ReRAM* CELL FABRICATION

FIGURE 5.1g, h shows the top view electric microscope image of the fabricated device, with the device structure schematic illustration in Fig. 5.1d, e. First, the $SiO_2$ substrate was sequentially cleaned by fuming nitric acid, the acetone, isopropyl alcohol, and deionized (*DI*) water in an ultrasonic oscillator. Then, the Metals films were deposited on the chemically cleaned $SiO_2$ substrate through electron beam evaporation in a vacuum chamber with the pressure of $10^{-8}$ torr. 5 nm Ti (the adhesion layer), 50 nm Pd (*BE*) were deposited by *ATC 2400* Sputtering System under vacuum level of $10^{-8}$ torr (this is the electron-beam evaporation and sputtering dual system that here we used the electron-beam evaporation system for metals deposition); 5 nm or 10 nm $HfO_2$ (the oxide layer) was deposited by Sputter (Alliance Concepts system) under vacuum level of $10^{-7}$ torr; 5 nm or 10 nm Ti (interface layer) and 50nm Pd (*TE*), which were deposited by *ATC 2400* Sputtering System under vacuum level of $10^{-8}$ torr. By lifting off technique, the nodes with dimensions are 5 µm × 5 µm are fabricated by negative photoresist (*AZnLof@2020*) and positive photoresist (*S1805*) for photolithography (Heidelberg microMLA) (see the profiles in Fig. 5.2). The pads dimensions are 15 µm × 15 µm. The *BE* via was etched by $CHF_3$/Ar gases (Sentech Etchlab 200) for 270s. A $HfO_{2-x}$ layer was formed between the $HfO_2$ and Ti by the spontaneous "oxygen grabbing" reaction after the Ti/Pd or Ti/Pt deposition process [116–119]. The optical microscopy image of the device was obtained by an Optical Microscope (Olympus bx51) in 50×. Scanning Eletronic Microscope (Hitachi S-4800) was applied to observe the details of the nodes and cross-section.



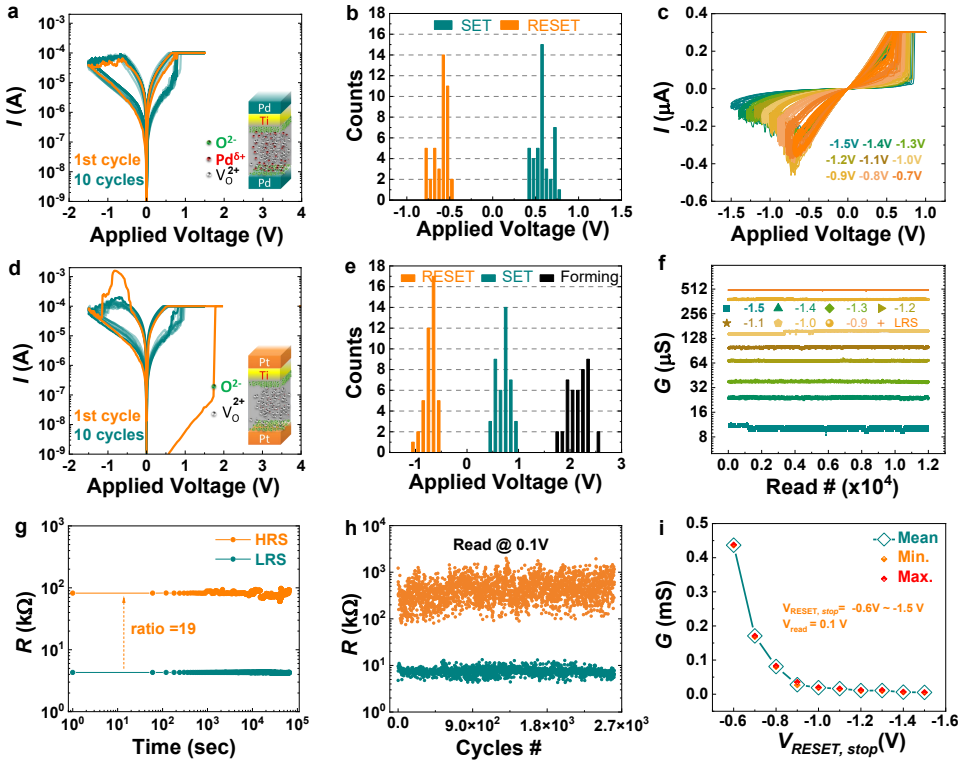

Figuur 5.2: Pd deposition on positive resist (a) and negative resist (c). (b), (d) are the cross section profile of as-fabricated different resists

### ELECTRICAL PERFORMANCE CHARACTERIZATION

To study electrical performance of the both structures cells, a probe station (*CASCADE*) equipped with a semiconductor analyser Keysight *B1500A* was utilized to conduct the electrical measurement for the *ReRAM* cell at room temperature (300 K), during which a bias voltage was applied to the *TE* while the *BE* was grounded. the double linear sweeping bias is from 0 V to 1.5 V with step of 100 mV for SET sweeping process and 0 V to −1.5 V with step of −100 mV for RESET sweeping process.

### *HRSTEM* CHARACTERIZATION

In our study, we implemented a multifaceted STEM methodology combining *HAADF*, *EELS*, and *iDPC* techniques to achieve a comprehensive characterization of advanced materials. We began with *STEM-HAADF* imaging, which utilizes a high-angle annular dark-field detector to collect thermally diffuse scattered electrons. This method provides Z-contrast images where the intensity scales roughly as $Z^{1.7}$, allowing us to resolve atomic columns and differentiate heavy and light elements based on their scattering power. Precise alignment and calibration are performed with dwell times in the order of µs to achieve high resolution and reliable compositional mapping.




Figuur 5.3: **Electrical characterization of PdHT and PtHT memristors**. **a, c** I-V curves of forming-free PdHT and PtHT conventional memristors, respectively. The inset schematic diagrams are corresponding structures. **b, d** statistical analysis of $V_{SET}$ and $V_{RESET}$ for PdHT and PtHT, respectively. **e** Tuning resistance states by changing DC sweeping $V_{RESET}$ stop voltages. **f** Stability test of 8 different conductance states by 12000 readings at 0.1 V. **g** Retention test of resistance states for 13000 seconds. **h** Endurance test of PdHT by DC sweeping 2500 cycles and reading at 0.1 V. **i** Conductance change over stop voltage of $V_{RESET}$ including its mean, maximum, and minimum conductance values.

*STEM-EELS* was employed to complement the HAADF images by providing elemental and chemical state information through electron energy loss spectroscopy. *EELS* spectra, collected in parallel with HAADF imaging, enabled the identification of subtle variations in bonding and composition at the nanoscale.

To enhance the detection of light elements and obtain electrostatic potential maps, we applied the *iDPC* method. In this mode, the segmented DF4 detector is selected via the Velox *DPC/iDPC* control panel, and the resulting differential signals are integrated. This process improves image contrast and reduces sensitivity to defocus and thickness variations, yielding an image where contrast is roughly proportional to atomic number.

### *RBS* CHARACTERIZATION

*RBS* was employed to determine the elemental depth profiles of the layers in the forming-free PdHT devices. *RBS* measurements were conducted using a *Kobe Steel HRBS-V500* system, with a $He^+$ ion beam accelerated at 400KeV. The beam was directed onto the sample at an incidence angle of 45°, and backscattered ions were detected at a scattering angle of 107.5°. A 512 channel detector with an energy resolution of 2 keV was used. The *RBS* data was analyzed using the Kobe Steel AnalysisIB software. The measurements revealed the Pd ratio in the $HfO_{2-x}$ layer, confirming the intermixing of Pd-O-Hf. This configuration is critical for reducing the oxygen diffusion barrier and facilitating the forming-free behavior.

### ACTIVATION ENERGY MEASUREMENT

The activation energy of the forming-free $HfO_2$-based *ReRAM* devices was measured using a Keithley *B1500A* Semiconductor Parameter Analyzer in conjunction with a thermionic heater. In this setup, the devices were mounted on a temperature-controlled stage that provided stable and uniform heating over a defined temperature range. To prevent oxidation and moisture interference during high-temperature measurements, a continuous flow of high-purity $N_2$ gas was maintained throughout the experiment. The *B1500A* recorded the current-voltage (I-V) characteristics at incremental temperature steps. The temperature-dependent resistivity was then derived from the I-V data, and an Arrhenius plot of $\ln(\rho)$ versus $1/T$ was constructed. The slope of the linear fit, when divided by Boltzmann's constant ($k_B$), yielded the activation energy. This method ensured precise thermal control and reproducible measurements, critical for understanding the conduction mechanisms in the devices.

### DEVICE CALIBRICATION AND POWER CONSUMPTION EVALUATION

The device model is developed based on the JART model, which accurately replicates the electrical behavior of the forming-free PdHT *ReRAM* devices. Calibration of the model is achieved by fitting the experimental *DC* I–V curves, obtained using a Keisight *B1500A*, with the simulation results. In our setup, both *DC* sweep and pulse-based measurement configurations are implemented using Cadence simulation tools. The *DC* sweep methodology enables the extraction of key parameters such as *SET*/*RESET* voltages and memory window, while pulse-based simulations reveal the dynamic switching characteristics and multi-resistance states under various programming schemes. This integrated simulation approach provides critical insights into the energy consumption during programming and reading operations, ensuring consistency between experimental and simulated results.

### DEVICE-BASED SYSTEM-LEVEL *SNN* SIMULATION

The synaptic weights of the two *SNNs* are quantized to integer values of 9 bits, which are represented by a combination of three multistate *ReRAM* devices with 8 states each, i.e., 3×3−bit *ReRAM* devices. In the presented experiments, we performed the training and inference for the two case studies and reported the write and read energy footprint of the PdHT and PtHT *ReRAM* devices in different scenarios. The

focus is kept on the energy impact of the crossbar arrays hosting the devices since it is not dependent on the actual realization of the periphery circuitry, allowing space for generalization no matter the actual implementation of the neuromorphic hardware accelerator. In all cases, we present a comparison between state-of-the-art PtHT *ReRAM* devices and the proposed PdHT ones.

***N-MNIST SNN:*** The *N-MNIST* dataset is a neuromorphic, that is, spiking, version of the MNIST dataset, which comprises images of handwritten arithmetic digits in grayscale format [120]. It consists of $7\times10^4$ sample images that are generated from the saccadic motion of a Dynamic Vision Sensor (*DVS*) in front of the original images in the MNIST dataset. The samples in the *N-MNIST* dataset are not static, i.e. they have a duration in time of 300 ms for each. The dataset is divided into a training set of $6.0\times10^4$ samples and a test set of $1.0\times10^4$ samples. The *SNN* architecture shown in Fig. 5.12a comprises 3 convolutional layers (SC1, SC2, and SC3) followed by 2 fully connected ones (SF4 and SF5). The classification accuracy on the test set is 94.56%, which is comparable to the performance of state-of-the-art level-based DNNs.

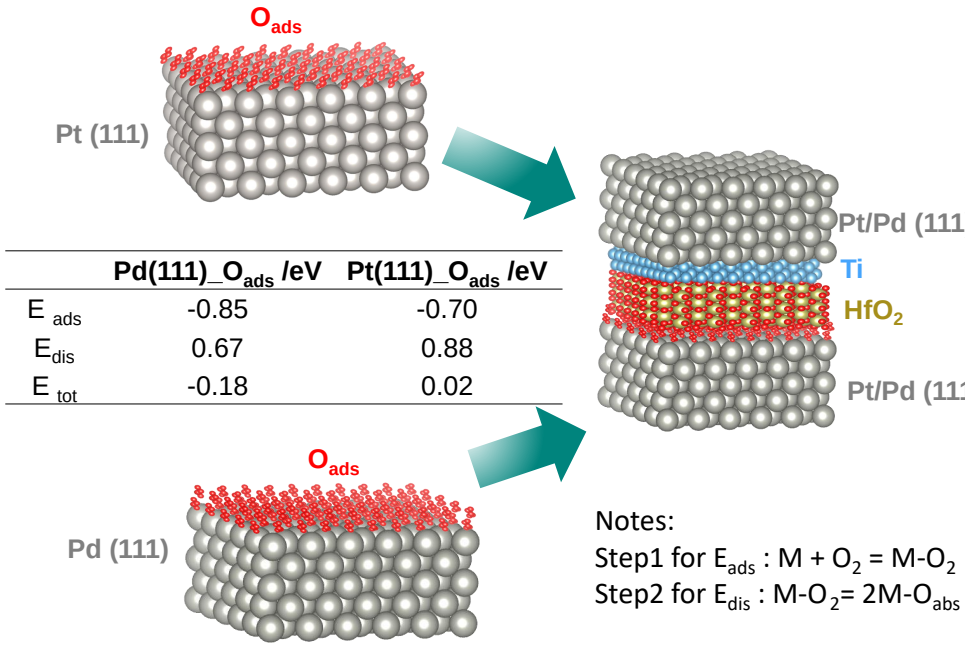
**Gesture *SNN*:** The IBM's *DVS128* Gesture dataset consists of 29 individuals performing 11 hand and arm gestures in front of a *DVS*, such as hand waving and air guitar, under 3 different lighting conditions [121]. Samples from the first 23 subjects are used for training, and samples from the last 6 subjects are used for testing. In total, the dataset comprises 1342 samples, each of which lasts about 6s, making the samples 20 times longer than those in *N-MNIST*. To speed up the neuromorphic simulations, we trimmed the length of the samples to about 1.5 s. The proposed *SNN* architecture consists of a pooling layer SP0 to reduce the input samples, 2 convolutional layers (SC1 and SC2) followed by a pooling layer each. The data coming from the last pooling layer is flattened and fed to 2 fully connected layers (SF3 and SF4). The network architecture is presented in Fig. 5.12b. The network performs with an accuracy of 85.61% in the test set, which is acceptable considering the shortened samples in the dataset and the shallower architecture compared to the originally proposed architecture in [121].

In both use cases, the *SNNs* are implemented with SRM spiking neurons, effectively capturing the temporal dynamics of spiking activity. Training uses a variant of the backpropagation algorithm, in which error is calculated on the probability of each neuron to change spike state, i.e., fire a spike if was in resting state or stop firing, in the next timing instance [122].

## 5.3. RESULTS AND DISCUSION

### DEVICE CHARACTERIZATION

We designed crossbar arrays of devices with 5 $\mu$m × 5 $\mu$m node size (see Method), as displayed in Figs. 5.1d, e. The successful realization of the designed layout and stack is validated by optical microscopy and High-Resolution Scanning Transmission Electron Microscopy (*HRSTEM*), as depicted in Figs. 5.1f, g. The distinct layers of the device stacks are clearly visible in the cross-sectional *HRSTEM* image, as shown in Fig. 5.1g, where the $HfO_{2-x}$ layer exhibits a thickness of approximately 5 nm, sandwiched by 5 nm layers of Ti and Pd electrodes.




| | Pd(111)_$O_{ads}$ /eV | Pt(111)_$O_{ads}$ /eV |
|---|---|---|
| $E_{ads}$ | -0.85 | -0.70 |
| $E_{dis}$ | 0.67 | 0.88 |
| $E_{tot}$ | -0.18 | 0.02 |

Figuur 5.4: Atomic schematics of oxygen absorbtion on Pd or Pt including absorption energy and dissociation barriers

Furthermore, *HRSTEM-EELS* imaging of PdHT (Fig. 5.1h) confirms the precise elemental distribution.

To examine the electrical properties of the fabricated devices, I–V characterizations were performed, as depicted in Fig. 5.3. Measurements on the electrical properties of the PdHT devices reveal three pivotal attributes for energy-efficient computing, including electroforming-free operation, low operating voltages, and tunable conductance. As illustrated in Fig. 5.3a, the PdHT devices manifest forming-free bipolar switching behavior. A statistical analysis of 42 randomly selected devices from the same die (Fig. 5.3b) reveals that the $V_{SET}$ predominantly centers around 0.56 V, while the $V_{RESET}$ is near −0.58 V. For comparison, similar characterizations were conducted on PtHT devices fabricated using an identical process except for choosing Pt as the top and bottom electrodes (*TE*/*BE*). As shown in Fig. 5.3d, these PtHT devices show bipolar switching, which is similar to PdHT devices, but they require an additional electroforming step (as illustrated in Fig. 5.3e), a result consistent with previous reports [111]. The Device-to-Device variability (*D2D*) test (Fig. 5.3e) confirms that PtHT devices require an electroforming voltage of approximately 2.3 V. Moreover, the statistical analysis presented in Table 5.2 confirms a significant alleviation in variability and a reduction in both $V_{SET}$ and $V_{RESET}$ for the PdHT structure compared to both the PtHT devices and the state-of-the-art reports [35, 37, 38], as detailed in Table 5.1. Table 5.1 presents a comprehensive comparison of state-of-the-art forming-free ReRAM devices, encompassing both

non-$HfO_2$ and $HfO_2$-based systems. This analysis reveals that these devices typically exhibit one or more inherent limitations: they require elevated SET/RESET voltages, offer a constrained range of resistance states, or depend on specialized treatments to eliminate the conventional electroforming process. In contrast, our work successfully overcomes these issues, providing a more efficient and robust solution.

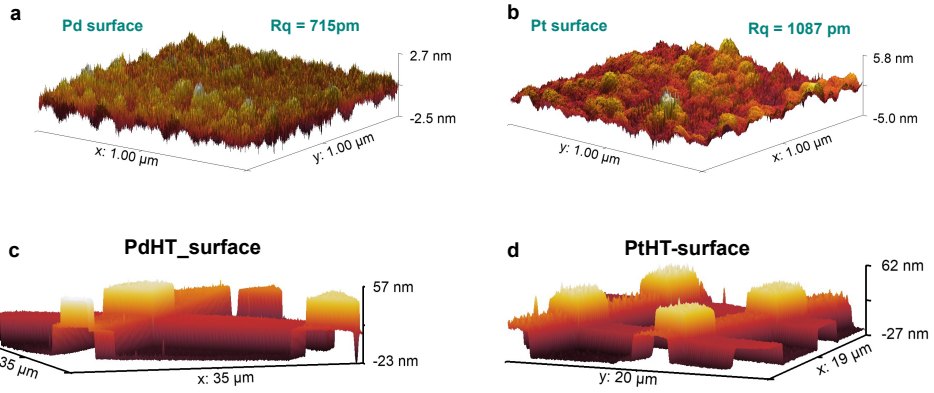
Tabel 5.2: Operating voltages and variation comparison of PtHT and PdHT devices.

| **Structures** | **C2C** | **D2D** | **$V_{SET}$** | **$V_{RESET}$** |
|---|---|---|---|---|
| Pt/Ti/$HfO_2$/Pt | 11.0% | 19.2% | 0.68 V | -0.72 V |
| Pd/Ti/$HfO_2$/Pd | 7.7% | 16.8% | 0.56 V | -0.58 V |

Furthermore, the reliability of the PdHT devices was also studied and certified by further electrical measurements. The multibit capability of the PdHT devices was evaluated using a linear voltage sweeping mode. As shown in Fig. 5.3c, the resistance is continuously modulated by varying the $V_{RESET}$, thereby precisely controlling the rupture degree of the *CF*. To ascertain the stability of the distinct resistance states, a *READ*-number dependent stability test was conducted (Fig. 5.3f), which demonstrates analog state functionality that permits the storage of eight tunable multi-bit weights within a single memory cell. Furthermore, retention (Fig. 5.3g) and endurance (Fig. 5.3h) assessments reveal the stable resistance states between *LRS* and High Resistance States (*HRS*), with outstanding non-volatility, where the resistance states remain stable for over $4.5 \times 10^4$ seconds, and the device can endure $2.5 \times 10^3$ cycles without noticeable performance degradation, though the performance is not presenting its optimum pattern (higher temperature of the retention test and higher number of cycles for the endurance test are needed for further investigation). Additionally, to gain deeper insight into the tunable resistance states, we conducted a comprehensive analysis of the conductance variation versus $V_{RESET,\ stop}$, as depicted in Fig. 5.3i. The results reveal an exponential gradual decline in conductance with increasing $V_{RESET,\ stop}$, indicating that the extent of *CF* rupture can be modulated by adjusting the amplitude of $V_{RESET,\ stop}$. This tunability may be further enhanced by controlling additional pulse parameters, such as pulse numbers, pulse amplitudes, and pulse width of $V_{RESET,stop}$.

### UNDERSTANDING THE FORMING-FREE BEHAVIOUR

We propose a possible explanation for the forming-free behavior (see Fig. 5.7a) based on a suite of comprehensive characterizations, including the Pd/$HfO_{2-x}$ interface imaging by *HRSTEM-iDPC* (*iDPC* stands for integrated Differential Phase Contrast.) (Fig. 5.7b), the surface profiles of Pd and Pt thin films obtained via Atomic Force Microscopy (*AFM*) (Fig. 5.5a, b), the Arrhenius activation energy measurements for both PdHT and PtHT memristors using a semiconductor analyzer equipped with a programmable thermoheater and an $N_2$-fluxed probe station (Fig. 5.7c), and the elemental distribution profiles acquired from Rutherford Backscattering Spectroscopy (*RBS*) (Fig. 5.7d). The activation energies for the pristine PtHT and PdHT devices are




Figuur 5.5: AFM surface analysis of PdHT device (left) and PtHT device (right)

compared in Fig. 5.7c via the Arrhenius equation:

$$\rho(T) = \rho_0 \exp\left(\frac{E_a}{k_B T}\right) \tag{5.1}$$

where $\rho(T)$ denotes the resistivity, $T$ is the absolute temperature in Kelvin, $\rho_0$ is the pre-exponential factor, $E_a$ represents the activation energy, and $k_B$ is the Boltzmann constant. As illustrated in Fig. 5.7c, the activation energy exhibits two distinct regimes: a High-Temperature (*HT*) regime (400 K to 470 K) characterized by larger activation energies (denoted as $E_{a2}$), and a Low-Temperature (*LT*) regime (300 K to 400 K) marked by lower activation energies (denoted as $E_{a1}$). Notably, the pristine PdHT devices exhibit activation energies of $E_{a1} = 0.0551\,\text{eV}$ and $E_{a2} = 0.220\,\text{eV}$, which are significantly lower than those of the PtHT devices ($E_{a1} = 0.124\,\text{eV}$ and $E_{a2} = 0.502\,\text{eV}$). The $E_{a2}$ of PtHT is attributed to the diffusion barrier associated with the migration of doubly positively charged oxygen vacancies, in accordance with previous reports [110, 123]. In contrast, the $E_{a2}$ of PdHT arises from the diffusion barrier of doubly negatively charged interstitial oxygen ions, which is in good agreement with *ab initio* calculations [110, 124]. It is noteworthy that the Frenkel-Pair (*FP*) dissociation process in $HfO_2$ typically requires activation energies on the order of 5.2 eV [110, 125]. Furthermore, if the *FP* comprises a doubly positively charged pair ($V_O^{2+}$ + $O^{2-}$), its formation energy is 5.8 eV [126], a value closely related to the electroforming process as reported previously [110, 127], implying that *FP* dissociation is highly unlikely to occur under both *LT* and *HT* conditions without an applied voltage bias.

In PtHT devices, the presence of $V_O^{2+}$ defects is likely a consequence of intrinsic defects introduced during the sputter deposition of amorphous $HfO_{2-x}$ [116]. Conversely, in PdHT devices, conduction is predominantly mediated by $O_i^{2-}$ defects. This behavior is likely due to the incorporation of Pd in $HfO_{2-x}$ at the $Pd/HfO_{2-x}$ interface and within the insulating $HfO_{2-x}$ layer, wherein Pd alters the defect landscape by donating electrons, thereby stabilizing oxygen vacancies as neutral ($V_O^0$) or fourfold-coordinated singly positively charged ($V_O^+$) defects, and forming $V_O^{q+}$–$O_i^{2-}$

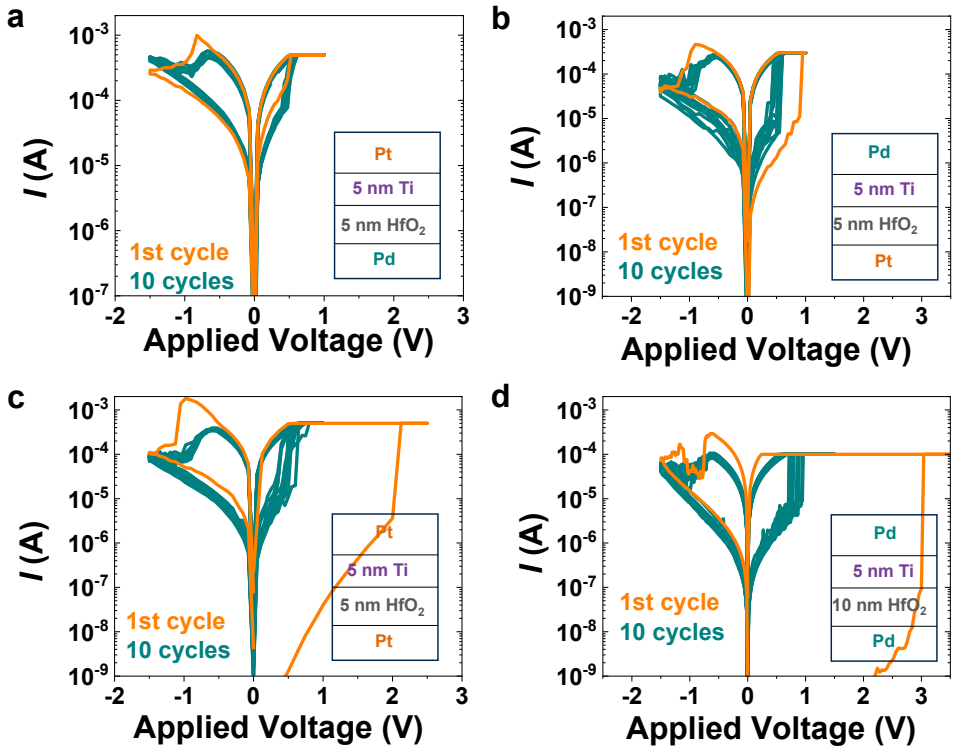
(q = 0, 1) pairs , serving as traps for electrons hoppting from hafnia conduction band [126]. This interpretation is further corroborated by additional activation energy measurements on the *HRS* state of both PdHT and PtHT memristors (Fig. 5.9), where it demonstrates that after electroforming (i.e., completion of *FP* dissociation) in PtHT devices, the dominant conduction exhibits an activation energy of approximately 0.22 eV, similar to that observed in PdHT devices. Alternatively, the observed behavior may also be ascribed to the diffusion of $V_O^{2+}$ defects among different coordinated sites; previous reports indicate that the diffusion barrier for $V_O^{2+}$ migration among fourfold coordinated sites is approximately 0.23 eV [123], suggesting that the formation pathways of $V_O^{2+}$ in $HfO_{2-x}$ differ between PtHT and PdHT devices due to the presence of Pd.

The incorporation of Pd is hypothesized to supply electrons that effectively reduce the diffusion barriers associated with oxygen vacancies or ions in the Pd–O–Hf configuration relative to the conventional Hf–O configuration in PtHT memristors. This reduction in barrier height facilitates the forming process [110] by effectively providing electrons to defective sites, as illustrated in Fig. 5.7a. Notably, this phenomenon predominantly occurs at the Pd/$HfO_{2-x}$ interface, where the insulating $HfO_{2-x}$ layer is thinned, thereby enhancing electron tunneling. To substantiate this hypothesis, *RBS* measurements were conducted. As depicted in Fig. 5.7d, the $HfO_{2-x}$ layer in the PdHT memristors exhibits a higher Pd concentration compared to the Pt concentration in PtHT devices. Remarkably, even after annealing the PdHT and PtHT devices at 573 K for 5 minutes, the Pd atomic ratio within the $HfO_{2-x}$ layer remains higher than that of Pt in the PtHT devices, as further illustrated in Fig. 5.13a, This observation is likely attributable to the higher propensity for Pd migration or diffusion into the $HfO_{2-x}$ layer, owing to lower migration barrier of Pd compared with other metals such as Li, Cu, Ag, Pt, and Au [128].

To confirm the presence of the Pd–O–Hf configuration, we employed *HRSTEM-iDPC* to examine the Pd/$HfO_{2-x}$ interface. As shown in Fig. 5.7b, the images reveal that the Pd-Hf atoms are clearly observed above Pd electrodes taken at the [011] incidence, showing a region where crystallized $HfO_{2-x}$ (evidenced by small, moderately bright atoms) intermixed with Pd (characterized by larger, highly bright atoms), while the central portion of the $HfO_{2-x}$ layer remains amorphous phase.

Meanwhile, *HRSTEM-iDPC* observations reveal that the Pd/$HfO_2$interface (Fig. 5.7b) maintains an optimal roughness level, a finding that is corroborated by AFM measurements (Fig. 5.5a, b). Statistical analysis shows that the Pd film exhibits a roughness of 715 pm, markedly lower than the 1087 pm measured for the Pt film, implying that the Pd surface is less prone to spike formation and thereby reduces potential measurement errors in *RBS* analysis. Furthermore, the *STEM-EELS* line profiles confirm the anticipated elemental composition in *LRS* (Fig. 5.13b), with the distinct elemental layers aligning precisely with the design illustrated in Fig. 5.1d. Notably, the presence of Pd-Hf at the interface of Pd/$HfO_{2-x}$ (*LRS*) verifies Pd-O-Hf configuration at the interface and in the $HfO_{2-x}$ layer, a possible consequence of Pd lower diffusion barrier and its higher propensity for $PdO_x$ formation [128, 129].

Apart from the aforementioned observations and analyses, we also investigated the band structures and Schottky barriers of the PdHT and PtHT memristors to




Figuur 5.6: Different combination of Pt and Pd as bottom/top electrodes and their I-V curve with 5 nm $HfO_{2-x}$ **a, b, c** and 5 nm of $HfO_{2-x}$ with Pd electrodes for top and bottom electrodes **d**.

ascertain whether the forming-free behavior arises from differences in Schottky barrier heights. From previous studies [130–133], it is well established that oxygen can chemisorb on noble metals such as Pt and Pd at room temperature, forming a stable layer of atomic oxygen or surface oxides. This phenomenon occurs due to the strong interaction between oxygen and the d-electrons of these metals under ambient conditions. In Fig. 5.4, we summarize the oxygen chemisorption energies on Pd(111) and Pt(111), along with the oxygen dissociation barriers, from which it is evident that Pd(111) exhibits a higher propensity to form Pd(111)-$O_{abs}$ than Pt(111). Such behavior significantly influences the work functions of Pd and Pt, as reported in previous studies [131, 132], where the effective work functions of Pt and Pd are in the ranges of 6.00–6.10 eV and 6.02–6.37 eV, respectively. Moreover, oxygen adsorption affects the work function differently: for Pt, it increases by +0.35–0.45 eV, whereas for Pd, the increase is +0.90–1.25 eV, a disparity attributed to the distinct dipole moments in Pt–O and Pd–O bonding. This differential impact is corroborated by the observation of oxygen ions via the *STEM-iDPC* technique at the Pd lattice in the Pd/$HfO_{2-x}$ interface (Fig. 5.8). In STEM as shown in Fig. 5.8, especially when combined with *HAADF* imaging, it serves as a complementary imaging mode. While

*HAADF* primarily provides Z-contrast (sensitive to atomic number, thus highlighting heavier elements), *iDPC* collects the differential phase contrast signals—often using a segmented or pixelated detector—and then integrates them to reconstruct the projected electrostatic potential of the sample. This integration allows for enhanced contrast of light elements and detailed structural information that might be missed by *HAADF* alone. From this image, we can observe that at the interface of Pd/$HfO_{2-x}$ oxygen is identified among Hf atoms and on Pd atoms.

Furthermore, the Charge Neutrality Level (*CNL*) for $HfO_{2-x}$ in both the Pd/$HfO_{2-x}$ and Pt/$HfO_{2-x}$ systems has been determined experimentally and theoretically, with the extracted *CNL* ($E_{CNL}$,$HfO_{2-x}$) reported as 4.36 eV [134]. By applying the simplified theory of Schottky barriers via Metal-Induced Gap States (MIGS), which attribute Schottky Barrier Height (SBH) pinning to a finite density of metal Fermi energy [135, 136], we use the relation:

$$\phi_{Bn} = S\left[\phi_M - \chi_{HfO_2}\right] + (1-S)\left[E_{CNL} - E_{CBM}\right] \tag{5.2}$$

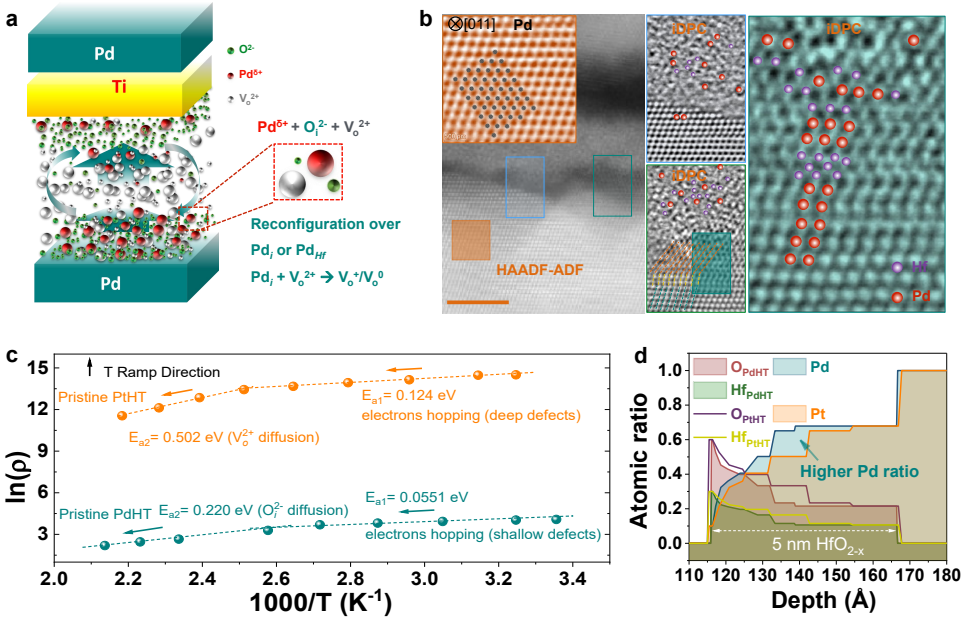
where $\phi_{Bn}$ is the Schottky barrier height, $\phi_M$ represents the effective work function of the metal, $S$ is the pinning factor at the metal/$HfO_{2-x}$ interface, $E_{CNL}$ is the charge neutrality level of the dielectric layer, $\chi_{HfO_2}$ is the electron affinity of $HfO_{2-x}$, and $E_{CBM}$ is the conduction band minimum. Based on this model (with $S = 0.52$, see ref. [135]), the calculated Schottky barrier heights for Pd/$HfO_{2-x}$ and Pt/$HfO_{2-x}$ are 4.17–4.22 eV and 4.18–4.37 eV, respectively, indicating that $\phi_{Bn}$ for Pd/$HfO_{2-x}$ is nearly equivalent to that of Pt/$HfO_{2-x}$. This result implies that the difference in activation energy $E_{a1}$ observed in Fig. 5.7c between the PdHT and PtHT memristors does not originate from disparities in Schottky barrier height. Consequently, this finding indirectly underscores the critical impact of the electrode materials on the forming-free behavior.

The influence of Pd on forming voltages is meticulously examined via the I–V characteristics of various devices (Figs. 5.6a, b). The critical role of Pd electrodes is underscored by their ability to markedly reduce or even eliminate the need for electroforming voltages, in stark contrast to the PtHT structure (Fig. 5.6c). Moreover, as the $HfO_2$ thickness increases, a distinct electroforming behavior emerges (Fig. 5.6d), likely due to the finite extent of the Pd–O–Hf configuration within $HfO_2$, which results in an exponential increase in resistance with a linear increase in thickness.

For clean Pt and Pd without exposure to oxygen, the interface energy of a polar O-terminated interface decreases as the metal work function increases in the transition metal series, as the interfacial M-O bond strength decreases as the metal work function increases, as illustrated in report[135].

### INVESTIGATION OF CONDUCTION MECHANISM

Conduction mechanism studies are frequently undertaken to elucidate the fundamental electrical conduction properties during I–V sweeps to pave the way for enhanced device reliability. In *ReRAM* devices, multiple conduction mechanisms [137, 138], such as Ohmic conduction, trap-mediated conduction, Trap-Assisted Tunneling (TAT), Space-Charge-Limited Conduction (*SCLC*), Poole–Frenkel emission, Schottky



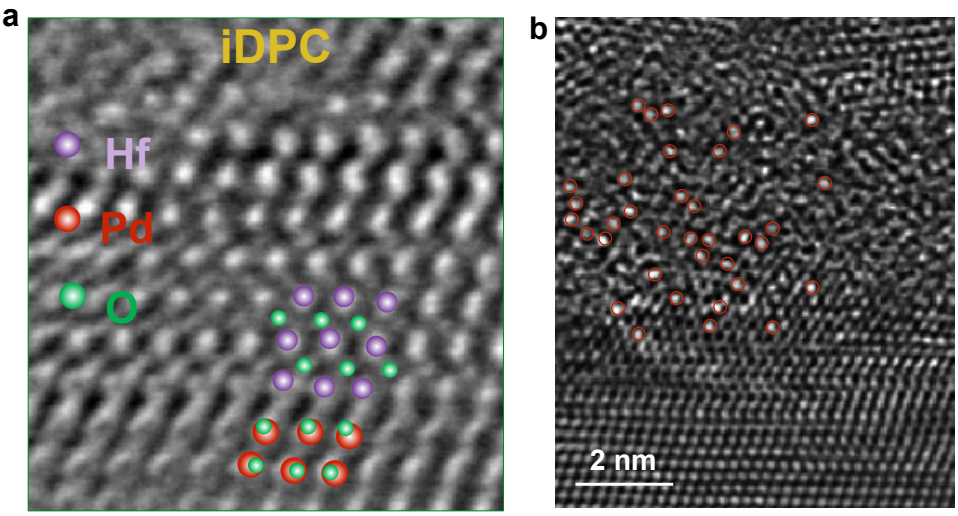

Figuur 5.7: | **Investigation of forming-free behavior a** schematic illustration of a forming-free mechanism. **b** *STEM-HAADF* and *iDPC* images of the PdHT device. The scale bar is 5 nm. **c** Activation energy study for PtHT (orange) and PdHT (dark cyan). **d** *RBS* elemental distribution line profile along PdHT cross section.

emission, and valence change-based mechanisms, collectively dictate the overall device behavior. The predominant conduction mechanism is inherently determined by factors including material selection, device architecture, and operating conditions.

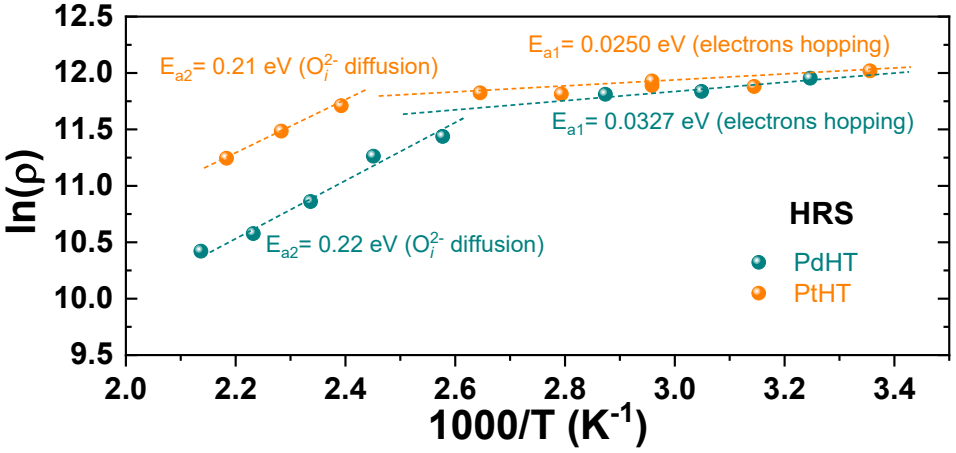
In particular, *SCLC* is a prominent mechanism observed in *ReRAM* devices during high-voltage operation [138]. Here, the current is constrained by a space-charge-limited region in which charge carriers, injected from the electrodes, traverse localized trap states within the resistive switching material. The *SCLC* behavior is critically influenced by the trap density and trap energy levels, which govern the mobility and transport of the charge carriers.

A trap-controlled *SCLC* regime can be segmented into two distinct portions: the trap-unfilled (trap-limited) regime and the trap-filled (trap-free) regime. In the Mott-Gurney region—characterized by the power law $I \propto V^2$—a steep increase in current is observed at high electric fields. This behavior is typically identified by an initial Ohmic conduction at low fields, followed by a transition to a power-law dependence as electrode-injected electrons surpass the equilibrium concentration. Consequently, *SCLC* conduction is more likely when the electrode contact exhibits high carrier injection efficiency [138].

The *SCLC* conduction mechanism is quantitatively described by the Mott-Gurney law (Eq. 5.3), which relates the current density $J$ to the applied voltage $V$ and




Figuur 5.8: **Oxygen ions observation using *STEM-HAADF-ADF-iDPC*** iDPC stands for integrated Differential Phase Contrast. In STEM, especially when combined with HAADF imaging, it serves as a complementary imaging mode. While HAADF primarily provides Z-contrast (sensitive to atomic number, thus highlighting heavier elements), iDPC collects the differential phase contrast signals—often using a segmented or pixelated detector—and then integrates them to reconstruct the projected electrostatic potential of the sample. This integration allows for enhanced contrast of light elements and detailed structural information that might be missed by HAADF alone. from this image, we can observe that at the interface of Pd/$HfO_{2-x}$ oxygen is identified among Hf atoms and on Pd atoms




Figuur 5.9: **Activation energy comparison for PdHT and PtHT over high resistance states (HRS)**

material properties:

$$J_{Mott-Gurney} = \frac{4}{9}\epsilon_0 \mu \frac{V^2}{d^3} \tag{5.3}$$

Here, $\epsilon_0$ denotes the dielectric permittivity, $\mu$ is the carrier mobility, $d$ represents the film thickness, and $V$ is the applied voltage.

The slope ($k$) of the log–log I–V curve evolves from approximately 1.0 (indicative of trap-limited Ohmic conduction) through 1.5 (reflecting the trap-unfilled transitional region) to nearly 2.0 (characteristic of trap-free *SCLC*) for both *HRS*and *LRS* as observed in Figs. 5.10a and d for PdHT devices under a 0.1 mA compliance current. On the contrary, when the compliance current is increased to 0.5 mA, the slope $k$ for *HRS* (as shown in Figs. 5.10b and e) is dominated by Ohmic conduction in both *HRS* and *LRS*, a consequence of reduced interface state density at the Pd/$HfO_{2-x}$ interface owing to filament formation. Conversely, PtHT devices exhibit Ohmic conduction under a 0.1 mA compliance current for both *HRS* and *LRS* (Figs. 5.10c and f).

Tabel 5.3: Energy consumption (PdHT). Measured parameters for various conductance states (G1–G8).

| | | | | | | | Averaged Power and Energy | | |
|---|---|---|---|---|---|---|---|---|---|
| **Conductance** | **R_states** | **$V_{reset\_stop}$ (V)** | **$V_{read}$ (V)** | **$Power_{read}$ (W)** | **$t_{read}$ (s)** | **$E_{read}$ (J)** | **$Power_{Program}$ (W)** | **$E_{programming}$ (J)** | **$E_{tot}$ (J)** |
| G1 | HRS1 | -1.5 | 0.1 | 7.48e-07 | 1.00e-05 | 7.48e-12 | 4.20e-04 | 4.20e-09 | 4.21e-09 |
| G2 | HRS2 | -1.4 | 0.1 | 9.15e-07 | 1.00e-05 | 9.15e-12 | 3.96e-04 | 3.96e-09 | 3.97e-09 |
| G3 | HRS3 | -1.3 | 0.1 | 1.14e-07 | 1.00e-05 | 1.14e-12 | 4.80e-04 | 4.80e-09 | 4.80e-09 |
| G4 | HRS4 | -1.2 | 0.1 | 1.44e-06 | 1.00e-05 | 1.44e-11 | 5.72e-04 | 5.72e-09 | 5.73e-09 |
| G5 | HRS5 | -1.1 | 0.1 | 1.88e-06 | 1.00e-05 | 1.88e-11 | 6.30e-04 | 6.30e-09 | 6.32e-09 |
| G6 | HRS6 | -1.0 | 0.1 | 2.57e-06 | 1.00e-05 | 2.57e-11 | 7.06e-04 | 7.06e-09 | 7.09e-09 |
| G7 | HRS7 | -0.9 | 0.1 | 3.80e-06 | 1.00e-05 | 3.80e-11 | 7.76e-04 | 7.76e-09 | 7.80e-09 |
| G8 | LRS | -0.5 | 0.1 | 5.09e-06 | 1.00e-05 | 5.09e-11 | 3.97e-04 | 3.97e-09 | 4.02e-09 |

These results indicate that at the Pd/$HfO_{2-x}$ interface in PdHT devices, shallow defect states are preferentially formed, enabling electrons to enter the conduction band (see Figs. 5.10g and h). In contrast, PtHT devices tend to form deep defect states, wherein electrons tunnel into vacancy defect sites [138] (see Fig. 5.10i). In PdHT devices, the shallow defects likely arise from intrinsic oxygen deficiencies as well as Pd–O–Hf induced defect states, which facilitate electron drift toward the Pd electrode via tunneling. Meanwhile, in PtHT devices, electrons predominantly hop between deep defect sites generated by high oxygen vacancy concentrations under higher electric fields following the electroforming process.

### DEVICE CALIBRATION AND ELECTRICAL SIMULATION

To evaluate the energy consumption of the PdHT and PtHT devices, we calibrated the *DC* I–V curves of the real devices using the *JART* model [139]. In the case

Tabel 5.4: Energy consumption (PtHT). Measured parameters for various conductance states (G1–G8).

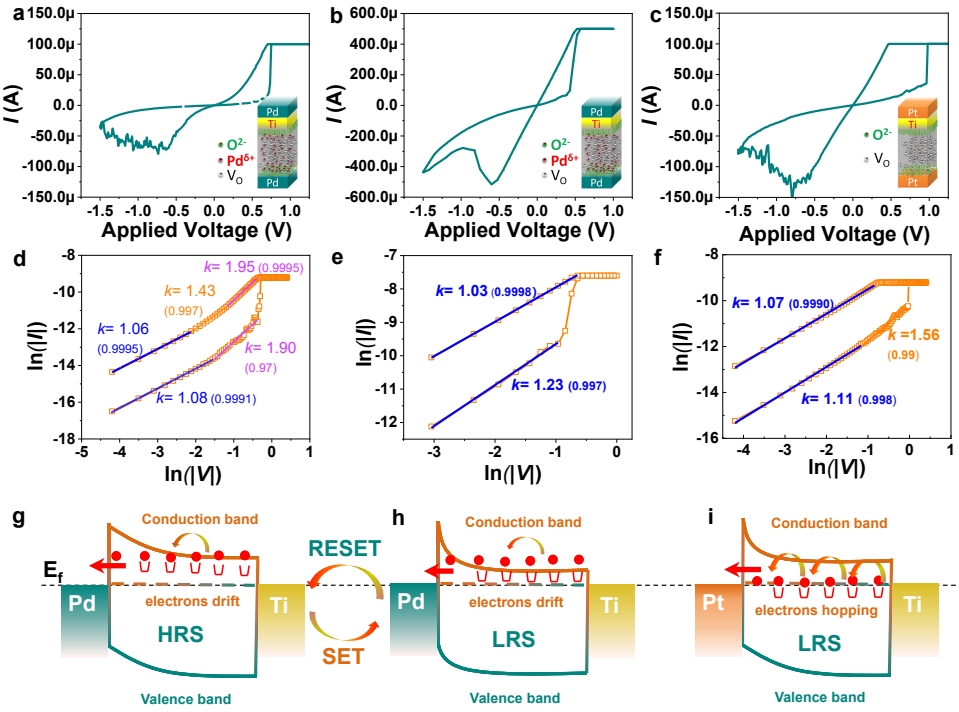
| | | | | | | | | Averaged Power and Energy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Conductance** | **R_states** | **$V_{reset\_stop}$ (V)** | **$V_{SET}$ (V)** | **$V_{read}$ (V)** | **$Power_{read}$ (W)** | **$t_{read}$ (s)** | **$E_{read}$ (J)** | **$Power_{Program}$ (W)** | **$E_{programming}$ (J)** | **$E_{tot}$ (J)** |
| G1 | HRS1 | -1.5 | 1.1 | 0.1 | 5.96e-06 | 1.00e-05 | 5.96e-11 | 8.92e-04 | 8.92e-09 | 8.98e-09 |
| G2 | HRS2 | -1.4 | 1.1 | 0.1 | 6.27e-06 | 1.00e-05 | 6.27e-11 | 9.44e-04 | 9.44e-09 | 9.51e-09 |
| G3 | HRS3 | -1.3 | 1.1 | 0.1 | 6.69e-06 | 1.00e-05 | 6.69e-11 | 1.00e-03 | 1.00e-08 | 1.01e-08 |
| G4 | HRS4 | -1.2 | 1.1 | 0.1 | 7.35e-06 | 1.00e-05 | 7.35e-11 | 1.07e-03 | 1.07e-08 | 1.08e-08 |
| G5 | HRS5 | -1.1 | 1.1 | 0.1 | 8.39e-06 | 1.00e-05 | 8.39e-11 | 1.12e-03 | 1.12e-08 | 1.13e-08 |
| G6 | HRS6 | -1.0 | 1.1 | 0.1 | 8.95e-06 | 1.00e-05 | 8.95e-11 | 1.07e-03 | 1.07e-08 | 1.07e-08 |
| G7 | HRS7 | -0.9 | 1.1 | 0.1 | 9.02e-06 | 1.00e-05 | 9.02e-11 | 9.78e-04 | 9.78e-09 | 9.87e-09 |
| G8 | LRS | -0.5 | 1.1 | 0.1 | 4.51e-06 | 1.00e-05 | 4.51e-11 | 5.88e-04 | 5.88e-09 | 5.93e-09 |



of PdHT, the simulated I–V performance exhibited excellent concordance with the experimental results, as shown in Fig. 5.11a. Regarding the multilevel resistance state performance, the rupture of the *CF* displayed a well-controlled, gradual transition upon tuning the $V_{RESET}$, as demonstrated in Fig. 5.11b, where over 10 distinguishable I–V curves were observed. Similarly, the incremental change in conductance with varying pulse amplitude for $V_{RESET}$ showed remarkable agreement between experimental observations and simulation results (Fig. 5.11c). Furthermore, distinct resistance states can be programmed using variable erase voltages ($V_{erase}$) with a 10 $\mu$s pulse duration, leading to the conclusion that 8 discernible resistance states can be established with corresponding programming schemes, as depicted in Fig. 5.11d. The detailed programming protocol is further illustrated in Fig. 5.11e, where each programmed resistance or conductance state is successfully achieved using a reading pulse of 10 $\mu$s at 100 mV. To assess the power consumption during programming, we integrated the instantaneous power $P(t) = I(t) \cdot V(t)$ over time, as shown in Fig. 5.11f; here, both RESET and SET energy consumptions contribute to the total programming energy, while the sum of *HRS* and *LRS* reading energies represents the total energy consumption during reading. The programming energy consumption for 8 distinct conductance states for both PdHT and PtHT devices is summarized in Table 5.3 and Table 5.4, respectively.

### SYSTEM-LEVEL ENERGY SIMULATION FOR NEUROMORPHIC APPLICATIONS

For the demonstration of the proposed PdHT *ReRAM* device in realistic neuromorphic applications, we designed two deep convolutional *SNNs* targeted at the classification of the *N-MNIST* [120] and IBM *DVS128* Gesture [121] datasets, respectively. These networks are implemented in Python using primitives from the open-source Spike LAYer Error Reassignment (SLAYER) [122] and PyTorch [140] frameworks, and they are trained with a variant of the backpropagation algorithm. Spiking neurons are




Figuur 5.10: **Investigation of conduction mechanism from I-V curves fitting for PtHT and PdHT. a** I-V curves of PdHT device with 0.1 mA compliance current. **b** 0.5mA compliance current along with their I-V curves fitting (d) and (e), respectively. (c), (f) present PtHT device IV curve and its IV curve fitting, respectively. **g, h** illusrate HRS and LRS band diagrams of PdHT and electrons drifting between shallow vacancies states, respectively. **i** shows the HRS band digram of PtHT and electrond hopping between deep defective sites.

modeled on the Spike Response Model (SRM), an advanced generalization of the ubiquitous Integrate-and-Fire (I&F) neuron model [141]. At the output, the class corresponding to the neuron that produced the highest number of spikes is selected as the winning class. For further methodological details, please refer to the Methods section.

To estimate the energy consumption at the system level, we perform a series of experiments and report the overall read and write energy measured collectively for all *ReRAM* devices needed to store the synaptic weights across each network. In the first set of experiments, we train the two *SNNs* in software and then we write the final synaptic weights directly onto the *ReRAM* devices. For simplicity and generalization purposes, there is no limit in the maximum number of employed devices. The first two bars in Figs. 5.12c and 5.12d report the write energy, $E_{write}$, per layer on a logarithmic scale for the *N-MNIST* and Gesture *SNNs*, respectively. As anticipated, the write energy scales proportionally with the number of synapses,

since a greater number of weights necessitates more devices. Notably, the PdHT devices achieve an overall reduction in write energy of approximately 43% for each *SNN*.

In the case of online learning, the synaptic weights are updated and written back on the *ReRAM* devices at every epoch of the training of each *SNN* and the overall write energy per epoch is illustrated in Figs. 5.12e and 5.12f for the two case studies, respectively. Initially, the synaptic weights are set to small random values, resulting in a peak write energy equal to the total write energy occurring by summing the corresponding bars for all layers in Figs. 5.12c and 5.12d. This is because in the first iteration almost all *ReRAM* devices need to be programmed to a new value. As training progresses, only those devices that require updates contribute to the write energy, leading to a gradual reduction that mirrors the convergence of the synaptic weights. Consequently, fewer updates are needed, thereby lowering the overall energy consumption.

Regarding the read energy $E_{read}$, during inference, represented by the rightmost two bars in Figs. 5.12c and 5.12d, the energy consumption per layer exhibits a pattern distinct from that of the write energy. In *SNNs*, each layer functions as a filter that reduces the number of spikes transmitted to subsequent layers, resulting in progressively sparser spiking activity. Therefore, the read energy depends not only on the number of synapses but also on the density of incoming spikes. For example, layer SF1 of the Gesture *SNN* consumes only 4 × more energy than its predecessor, despite having 228 × more synapses. Overall, both networks exhibit an approximate 73% reduction in read energy during the complete inference of their respective test sets. The average energy per spike during inference is summarized in Table 5.5 and demonstrates a comparable energy reduction when utilizing the proposed PdHT devices over PtHT ones.

Overall, the proposed PdHT ReRAM device can result in significant energy savings during both programming and inference of SNNs running on neuromorphic accelerators.

| | N-MNIST SNN | Gesture SNN |
|---|---|---|
| PtHT | 33.6 nJ | 65.7 nJ |
| PdHT | 9.4 nJ | 17.7 nJ |

Tabel 5.5: Average read energy per spike.

## 5.4. CONCLUSION OF THIS CHAPTER

IN this work, we report the successful CMOS-compatible fabrication of Pd/$HfO_{2-x}$ (PdHT) resistive memory devices, benchmarked against conventional Pt/$HfO_{2-x}$ (PtHT) structures. High-resolution imaging analyses, including SEM, STEM, and AFM, confirm well-defined device stacks with minimal interfacial roughness. Electrical characterization reveals *formation-free* switching at relatively low operating voltages, with $V_{SET}$ in the range of 0.56–0.75 V and $V_{RESET}$ spanning from −0.58 to −0.65 $V$.

Notably, the PdHT devices exhibit cell-to-cell (c2c) and device-to-device (d2d) variability comparable to their PtHT counterparts, indicating robust switching behavior and enhanced reliability. Both structures also demonstrate desirable endurance and retention properties, as confirmed by repeated cycling and long-term read tests.

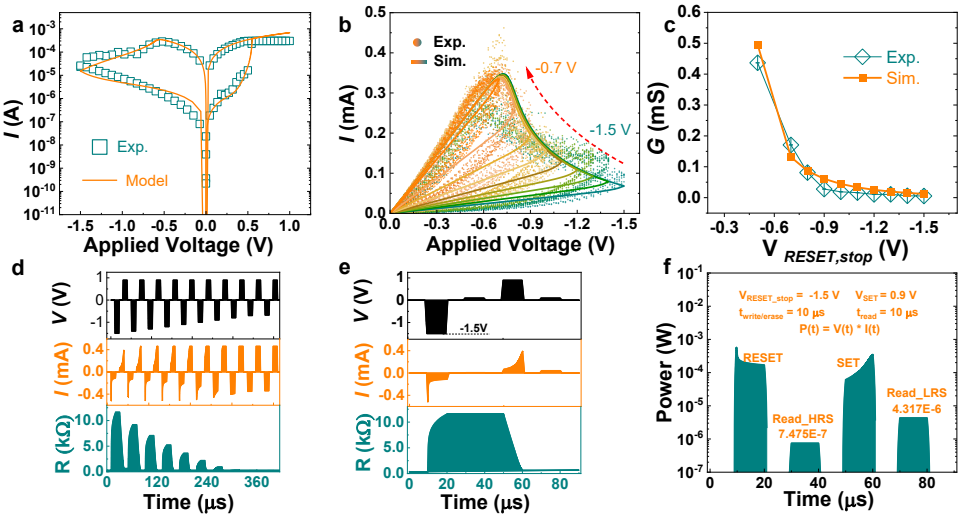
The absence of an electroforming step in PdHT devices is primarily attributed to the modified Pd/$HfO_{2-x}$ interface, where intrinsic Pd residence at the interface and in $HfO_{2-x}$ facilitates the formation of a Pd–O–Hf configuration. This interfacial structure effectively reduces the oxygen diffusion barrier, thus eliminating the need for a high-voltage forming process. Further evidence comes from activation energy measurements: PdHT devices exhibit a lower barrier (0.64 eV) relative to PtHT devices (1.44 eV), suggesting a more favorable distribution of oxygen vacancies and defect states. Microscopic and spectroscopic analyses (STEM-HAADF, iDPC, EELS, RBS) confirm a smoother interface with lower roughness and uniformly dispersed Pd within the $HfO_{2-x}$ matrix. In addition, the work function of Pd enhances oxygen chemisorption, suppressing deep-level defect formation and mitigating Frenkel-pair dissociation.



Current–voltage ($I$–$V$) analysis indicates that both PdHT and PtHT devices operate under a space-charge-limited conduction (SCLC) mechanism. At low bias, conduction follows an ohmic behavior governed by thermally generated free carriers. As the voltage increases, trap-mediated processes dominate; injected carriers fill shallow defect states, transitioning the conduction to a quadratic dependence on voltage ($I \propto V^2$), consistent with the Mott–Gurney law. In PdHT devices, the shallower defect states induced by the Pd–O–Hf interface enable smoother electron transport and lower variability compared to PtHT devices, which appear to harbor deeper defect levels.

The integration of these forming-free, multibit ReRAM devices into neuromorphic systems has been successfully demonstrated by employing them as synaptic elements in spiking neural networks (SNNs). Using three-device (3×3-bit) configurations to represent 9-bit quantized synaptic weights, the networks are implemented for real-world pattern recognition tasks on the N-MNIST and IBM's DVS128 Gesture datasets. The SNN architectures—developed using frameworks such as SLAYER and PyTorch—achieve competitive classification accuracies (94.6% for N-MNIST and 85.6% for gesture recognition). Furthermore, energy consumption analyses reveal that the PdHT devices yield substantial energy savings, with write and read operations reduced by approximately 43% and 73%, respectively.

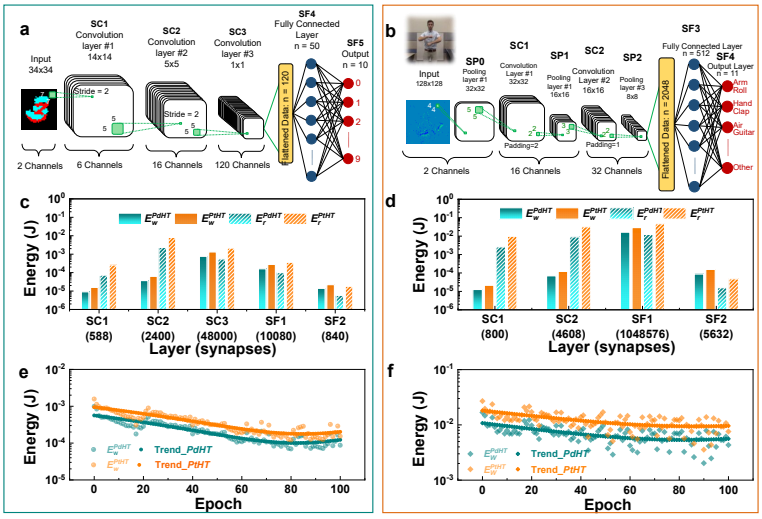
This energy efficiency, combined with stable multilevel conductance programming and robust endurance, makes the proposed ReRAM technology highly promising for next-generation, low-power neuromorphic computing applications.




Figuur 5.11: **Cadence simulation for power consumption evaluation a** Experimental DC I-V curve calibration with JART memristors model. **b** multistate testing on Cadence using DC sweeping by changing $V_{RESET,\ stop}$. **c** Comparison of conductance gradual change for experimental and simulation results. **d** 8 resistance states programming schemes.**e** Programming scheme of $V_{RESET,stop} = -1.5\,V$. **f** Power consumption evaluation for programming minimum conductance states.

Tabel 5.6: Write Energy: loading pre-trained nets for NMNIST SNN.

| | | Write Energy (J) | | Average Write Energy per Synapse (J) | |
|---|---|---|---|---|---|
| **Layer** | **Synapses** | **Pt** | **Pd** | **Pt** | **Pd** |
| SC1 | 588 | 1.57e-05 | 9.15e-06 | 2.67e-08 | 1.56e-08 |
| SC2 | 2400 | 6.31e-05 | 3.68e-05 | 2.63e-08 | 1.54e-08 |
| SC3 | 48000 | 1.34e-03 | 7.69e-04 | 2.78e-08 | 1.60e-08 |
| SF1 | 10080 | 2.79e-04 | 1.62e-04 | 2.77e-08 | 1.61e-08 |
| SF2 | 840 | 2.22e-05 | 1.40e-05 | 2.64e-08 | 1.67e-08 |
| Total | 61908 | 1.72e-03 | 9.91e-04 | 2.77e-08 | 1.60e-08 |




Figuur 5.12: **Hardware implementation of spiking neural network for N-MNIST and gesture recoginition a** schematic illustration of SNN structure for the N-MNIST dataset and its corresponding energy consumption comparison for PdHT and PtHT devcies in different layers **c**. **b** schematic illustration of SNN structure for gesture dataset and its corresponding energy consumption comparison for PdHT and PtHT devcies in different layers. **e** Energy consumption change dependent on epoch for the aforementioned SNN structure for N-MINST datasets. **f** Energy consumption change dependent on epoch for aforementioned SNN structure for gesture datasets.

Tabel 5.7: Write Energy: loading pre-trained nets for Gesture SNN.

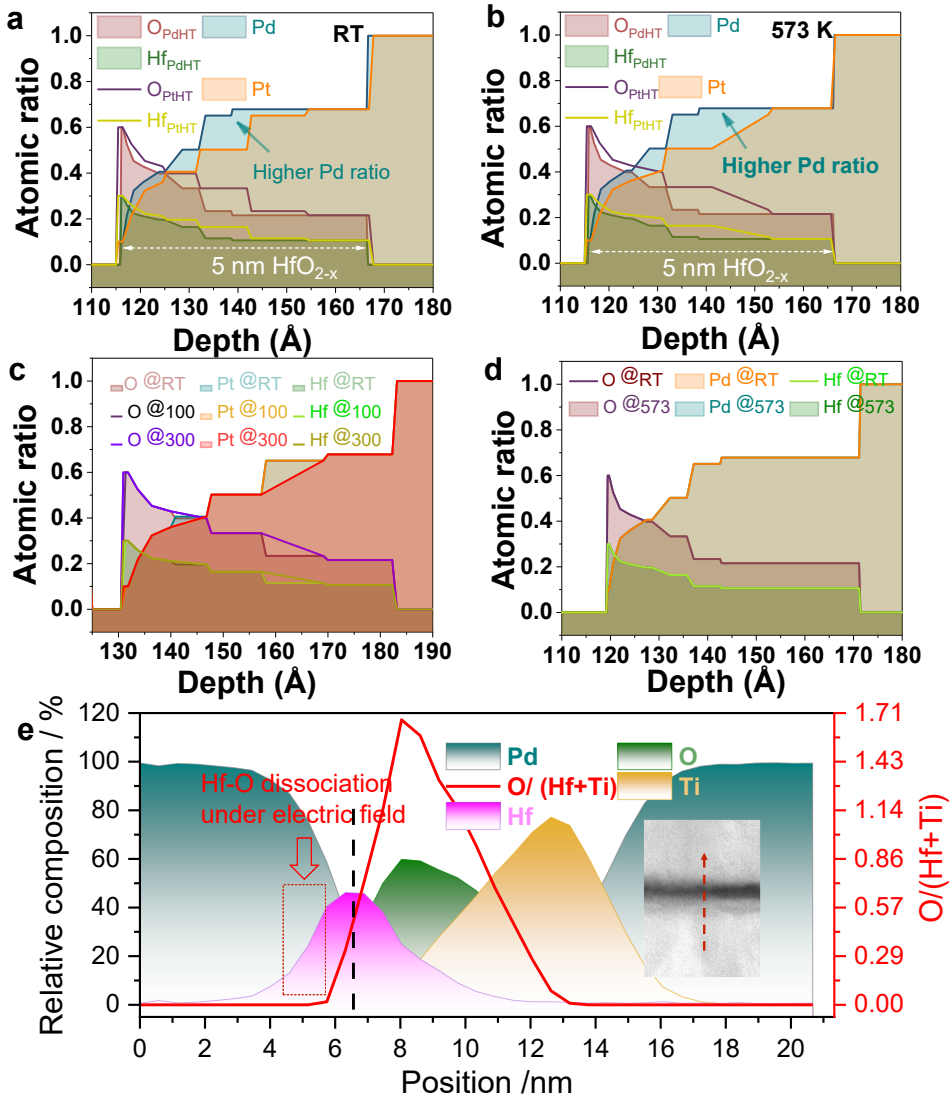
| | | **Write Energy (J)** | | **Average Write Energy per Synapse (J)** | |
|---|---|---|---|---|---|
| **Layer** | **Synapses** | **Pt** | **Pd** | **Pt** | **Pd** |
| SC1 | 800 | 2.15e-05 | 1.25e-05 | 2.69e-08 | 1.57e-08 |
| SC2 | 4608 | 1.21e-04 | 7.06e-05 | 2.63e-08 | 1.53e-08 |
| SF1 | 1048576 | 2.91e-02 | 1.66e-02 | 2.77e-08 | 1.58e-08 |
| SF2 | 5632 | 1.55e-04 | 8.86e-05 | 2.76e-08 | 1.57e-08 |
| Total | 1059616 | 2.94e-02 | 1.68e-02 | 2.77e-08 | 1.58e-08 |

Tabel 5.8: Read Energy: NMNIST SNN. Measured data for each layer (SC1, SC2, SC3, SF1, SF2) and total.

| | | Read Energy (J) | | Average Read Energy per Synapse (J) | |
|---|---|---|---|---|---|
| **Layer** | **Synapses** | **Pt** | **Pd** | **Pt** | **Pd** |
| SC1 | 588 | 2.61e-04 | 7.36e-05 | 4.43e-07 | 1.25e-07 |
| SC2 | 2400 | 8.04e-03 | 2.29e-03 | 3.35e-06 | 9.55e-07 |
| SC3 | 48000 | 2.17e-03 | 5.55e-04 | 4.52e-08 | 1.16e-08 |
| SF1 | 10080 | 3.65e-04 | 9.93e-05 | 3.62e-08 | 9.83e-09 |
| SF2 | 840 | 1.82e-05 | 5.85e-06 | 2.17e-08 | 6.96e-09 |
| Total | 61908 | 1.08e-02 | 3.03e-03 | 1.75e-07 | 4.89e-08 |

Tabel 5.9: Read Energy: Gesture SNN. Data for each layer (SC1, SC2, SF1, SF2) and a total row.

| | | Read Energy (J) | | Average Read Energy per Synapse (J) | |
|---|---|---|---|---|---|
| **Layer** | **Synapses** | **Pt** | **Pd** | **Pt** | **Pd** |
| SC1 | 800 | 9.93e-03 | 2.63e-03 | 1.24e-05 | 3.29e-06 |
| SC2 | 4608 | 3.23e-02 | 9.23e-03 | 7.01e-06 | 2.00e-06 |
| SF1 | 1048576 | 4.69e-02 | 1.23e-02 | 4.48e-08 | 1.17e-08 |
| SF2 | 5632 | 4.85e-05 | 1.58e-05 | 8.61e-09 | 2.81e-09 |
| Total | 1059616 | 8.92e-02 | 2.42e-02 | 8.42e-08 | 2.28e-08 |




Figuur 5.13: **a,b** Atomic ratio comparison of PtHT and PdHt under room temperature ("RT") and 573K. **c,d** Atomic ratio comparison of PtHT dependent on temperature and that of PdHT (the temperature unit is Kelvin). **e** STEM-EELS line profile of LRS state for PdHT, where it shows the Hf ions and oxygen ions dissociation, and Hf migrates into the Pd layer while oxygen ions migrate into Ti layer.

# 6

# CRYO-MEMRISTORS FOR QUBITS SPIN CONTROL

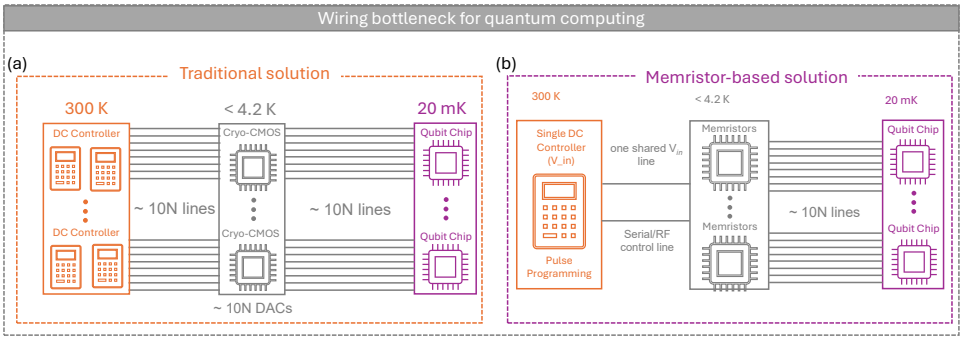
*We demonstrate interface-enhanced memristors (OxReRAM) tailored for cryogenic spin-qubit control. By engineering a sparse filament network, our devices achieve eight nonvolatile resistance levels with an ultra-low read noise rate of around 0.3 %. When embedded in a cryogenic gain stage with $R_L = 30\,\mathrm{k}\Omega$ and $V_{in} = 0.3\,\mathrm{V}$, it will deliver a ±1 V output range and sub-100-μV resolution using only six memristors per channel. This single-line biasing architecture will reduce wires, paving the way for large-scalce quantum processors. Meanwhile, we proposed a method to eliminate the wiring bottleneck though in-situ memristors-quantum dots devices integration.*

The contents of this chapter were based on the accepted 2025 IEDM paper [142].

## 6.1. INTRODUCTION OF THIS CHAPTER

SEMICONDUCTOR quantum dots (QDs) now exhibit gate fidelities exceeding 99 % and coherence times from milliseconds to seconds, establishing them as leading platforms for fault-tolerant quantum processors [143, 144]. However, scaling from tens to thousands of qubits demands tens to hundreds of independent, ultra-low-noise DC biases per qubit for controlling potentials without overloading the dilution refrigerator cooling power or wiring capacity [44, 145]. For nowadays standard approach, each qubit requires ~10 control electrodes, barrier, plunger, and screening gates, each routed through individual low-pass filters and cryogenic DACs at 4.2 K, then driven by high-speed links from 300 K controllers, as shown in Fig. 6.1(a). This $\mathcal{O}(N)$ wiring load imposes prohibitive heat leaks, cable bulk, and system complexity [145–147]. Malinowski *et al.* introduced the WISE architecture, embedding switch matrices on-chip to share waveform generators across many electrodes, reducing room-temperature DAC counts by an order of magnitude while still relying on multiple high-bandwidth feeds and analog filters [145]. A more radical solution leverages cryogenic nonvolatile memories: Mouny *et al.* demonstrated a memristor-based programmable gain amplifier (PGA) that replaces each DAC–filter pair with a single shared bias line and on-chip memristors crossbars at 4 K, collapsing feedthroughs to $\mathcal{O}(1)$ [44]. These works establish the feasibility of sub-µW, sub-µV-noise amplification at millikelvin temperatures. There, a common DC bias $V_{in}$ and a serial/RF interface program memristor conductances in the feedback of *Cryo-CMOS* operational amplifier (OpAmp), yielding up to ten independent, sub-100-µV-resolved outputs per channel (see Fig. 6.1(b)). Meanwhile, the problem





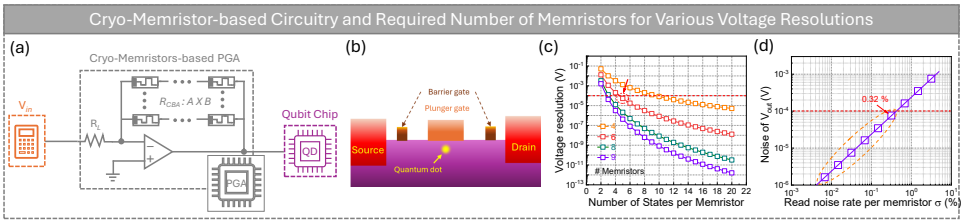

Figuur 6.1: Schematic comparison of control wiring for large-scale quantum processors. (a) Traditional solution: Each of $N$ qubits at 20 mK uses ~ $10N$ DC electrodes, routed through low-pass filters into ~ $10N$ *Cryo-CMOS* DACs (< 4.2 K), each driven by its own high-speed digital link from 300 K. This requires $\mathcal{O}(N)$ digital and coax lines, creating thermal load and cabling complexity. (b) Cryo-Memristor-based solution: A single shared bias line $V_{in}$ and one serial/RF link from 300 K feed on-chip memristor -based Programmable Gain Amplifiers (PGAs) at 4.2 K, which locally synthesize ~ $10N$ electrode voltages. Wiring scales as $\mathcal{O}(1)$, reducing heat load and enabling scalable qubit integration.

is that the read noise rate is around 1 % which is still not feasible for sub-100-μV step of gate voltage control. Complementary monolithic 65 nm CMOS studies have validated AD8605 amplifiers down to 1.2 K and $TiO_2$ memristor programming with 10 mV steps, enabling $\sim 10^6$ channels within a 1.5 W 4 K budget [21] while the problem is still there that the resolution is 10 mV which is still far from sub-100-μV precision. Nonetheless, key challenges remain [148]: achieving ultra-low programming read noise rate at 4.0 K and balancing the trade-offs among number of programming levels, read noise rate and output-voltage resolution. In this work, we propose, for the first time, modified interface-enhanced Pt/Ti/$HfO_2$/Pt memristors (M-PtHT) optimized for 4 K operation with eight-level programmability. Mounted on a *Cryo-PCB* test board, we evaluate resistance reliability, including retention and read noise rate. We demonstrate a read noise rate of around 0.3 %, and when integrated into a cryogenic PGA ($R_L = 30\,\text{k}\Omega$, $V_{in} = 0.3\,\text{V}$), it will guarantee a ±1 V range of output voltage with <100 μV resolution using minimum 6 memristors in corssbar arrays. These results confirm memristor-PGA technology as a compact, scalable approach for next-generation quantum processors.

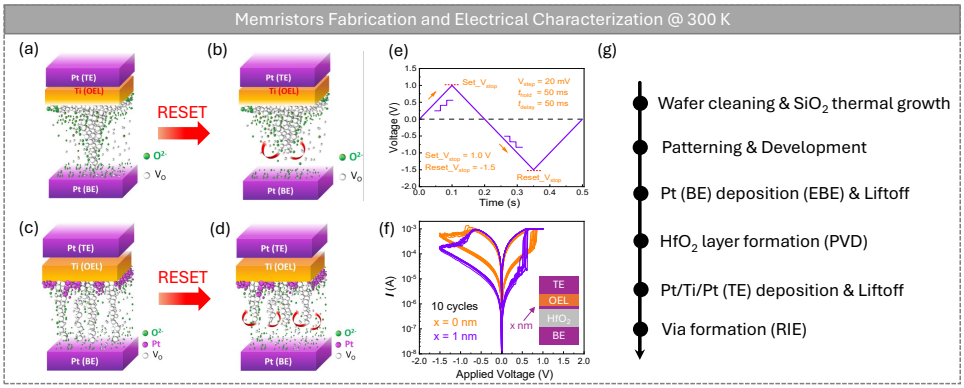
## 6.2. RESUTLS AND DISCUSSIONS

### 6.2.1. CRYO-MEMRISTOR-BASED CONTROL CIRCUITRY AND RESOLUTION SIMULATION






Figuur 6.2: Cryo-Memristor-based control circuitry for quantum dot devices and its performance. **(a)** Schematic configuration of the spin qubit control: a single room-temperature DC bias $V_{in}$ is applied through load resistor $R_L$ (30 kΩ) to on-chip memristor crossbar arrays (*CBA*) with A × B size in feedback around a *Cryo-CMOS* operational amplifiers (*OpAmp*), generating multiple individually tuned outputs for quantum-dot gates. **(b)** Cross-section schematic of a lateral quantum dot, showing barrier and plunger gates that require precise DC biases. **(c)** Simulated voltage resolution versus states per memristor for arrays of 4–9 devices, demonstrating sub-100-μV resolution with modest array size (e.g. 6 memristors with 5 states), given the $V_{in}$= 0.3 V, to make sure the $V_{out}$ is ranging from 0.1 V to 1.0 V. **(d)** The Monte Carlo simulation to show how single-device read-noise rates (0.003 %–3 %) translate into output-voltage noise $\sigma(V_{out})$ for a 6 × 5-state memristors PGA, where less than 0.32 % read noise rate per memristor is required to make sure the noise of $V_{out}$ is less than the $V_{out}$ resolution.




Figuur 6.3: Fabrication process and electrical characterization of the Pt/Ti/$HfO_2$/Pt (PtHT) memristor. **(a,b)** RESET mechanism: oxygen vacancies migrate away from the $HfO_2$/Pt interface under a negative voltage on top electrodes (TE) while bottom electrodes (BE) are grounded, rupturing the multiple conductive filaments. **(c,d)** Modified memristors PtHT (M-PtHT with 1 nm of innert interfacial metal layer): vacancy drift at the Ti/$HfO_2$ interface similarly disrupts the filament. **(e)** Bipolar voltage waveform used for set/reset cycling schemes. **(f)** Ten consecutive I–V loops demonstrating stable switching between low- and high-resistance states for PtHT and M-PtHT. **(g)** Device fabrication flowchart: EBE: Electron Beam Evaporator; PVD: Physical Vapor Deposition; RIE: Reactive Ions Etching.

Figure 6.2 illustrates the cryogenic memristor-*PGA* for quantum-dot biasing and its simulated resolution. In Fig. 6.2a, a single 0.3 V DC bias passes through $R_L = 30\,\text{k}\Omega$ into a *Cryo-CMOS OpAmp* whose feedback is an $N_m$-device memristor crossbar array (*CBA*). Programming each conductance $G_i$ sets:

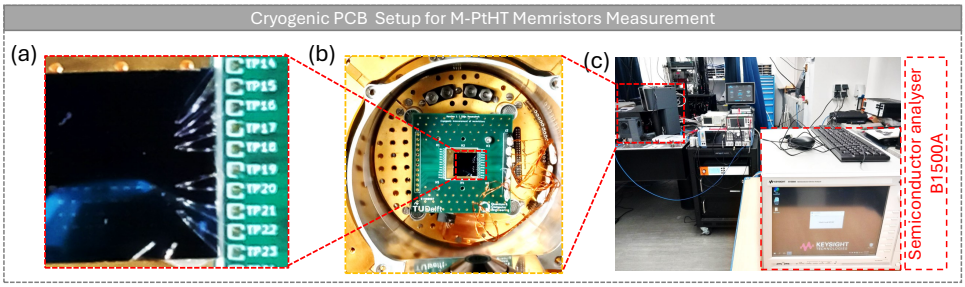
$$R_{\text{CBA}} = \left(\sum_i G_i\right)^{-1}, \quad V_{\text{out}} = V_{\text{in}} \frac{R_{\text{CBA}}}{R_L} \approx 1\ \text{V} \tag{6.1}$$

covering ±1 V without individual DACs. The QD cross-section (Fig. 6.2b) shows source/drain reservoirs, two barrier gates, and a plunger gate, all requiring sub-100-µV precision. Defining resolution $\delta V = (V_{\max} - V_{\min})/N_s^{N_m}$ with a 1 V range, Fig. 6.2c plots $\delta V$ versus states $N_s$ for $N_m = 4, 6, 8, 9$. Six-device arrays with five states yield $\delta V < 100\,\mu\text{V}$, demonstrating that small *CBAs* meet qubit-biasing needs while reducing wiring to a single bias line and programming link.

### 6.2.2. FABRICATION AND ELECTRICAL CHARACTERIZATION OF (M-)/PTHT MEMRISTORS

Figure 6.3 summarizes device structure, switching physics, test waveforms, I–V cycling, and fabrication flow. In PtHT devices (Fig. 6.3a,b), negative bias drives oxygen vacancies away from the $HfO_2$/Pt interface, rupturing filaments and switching

to a high-resistance state. The M-PtHT variant adds a 1 nm interfacial Pt layer (Fig. 6.3c,d), which suppresses stochastic vacancy drift at the Ti/$HfO_2$ interface, sharpening state separation and reducing variability. We apply a bipolar staircase waveform (20 mV steps, 50 ms hold/delay, ±1.0/–1.5 V stop voltages; Fig. 6.3(e) to program eight levels, then record ten consecutive I–V loops (Fig. 6.3(f). Both stacks exhibit stable LRS/HRS switching, with M-PtHT showing an expanded memory window. All devices are fabricated on $Si/SiO_2$ using PVD for $HfO_2$, and EBE for Ti adhesion and Pt layers, with optical inspection ensuring pad integrity (Fig. 6.3g). At 300 K, a Keysight B1500A applies a 3.5 V electroforming pulse followed by ten SET/RESET cycles (±1.0/–1.5 V) to assess variability. Multi-level programming uses variable RESET stops, and retention is measured under 0.2 V reads every 35 ms over 800 s to quantify drift and read noise rate. Cryogenic tests on wire-bonded chips mounted on a KiCad PCB board in a 4.0 K cryostat repeat the same protocols (Fig. 6.4), confirming device robustness under cycling and retention test.




Figuur 6.4: Cryogenic measurement setup for M-PtHT memristors: (a) M-PtHT memristor chip on test PCB; (b) Mounted on 4 K stage with wire bonds; (c) External Keysight B1500A semiconductor analyzer controlling and reading via cryostat feedthroughs.

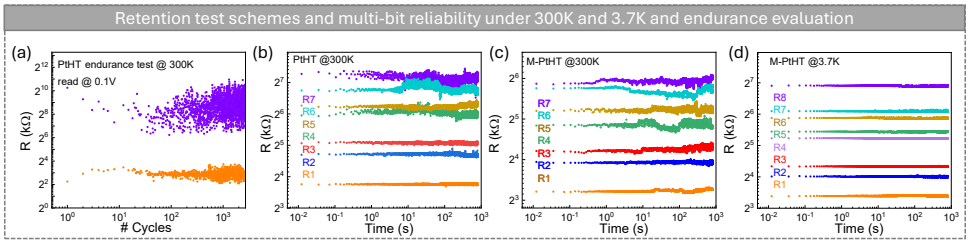
### 6.2.3. RELIABILITY UNDER 300 K AND 4.0 K

Figure 6.5 summarizes endurance cycling (a) and seven-level retention for both PtHT and M-PtHT memristors (b-d) at room temperature (300 K) and cryogenic (4.0 K) conditions. PtHT devices are subjected to over 3 $\times 10^3$ SET/RESET cycles with a read bias of 0.2 V. The LRS (orange) and HRS (purple) remain well separated, demonstrating robust switching and negligible degradation. After programming seven distinct resistance levels (R1–R7) at 300 K, each level is monitored under 0.2 V bias for 800 s. The levels remain drift-limited spread but few states collisions, especially for higher resistance states. Under cryogenic operation, M-PtHT devices exhibit even tighter level clustering, higher number of resistance states and alleviated drift and read noise over 800 s, confirming their suitability for long-term QD gate biasing with descent stability. Therefore, these results validate that the Pt-based devices deliver sufficient endurance cycle for resistive switching, multilevel retention, and cryogenic reliability required for programmable, on-chip DC-bias generation in

scalable quantum processors.

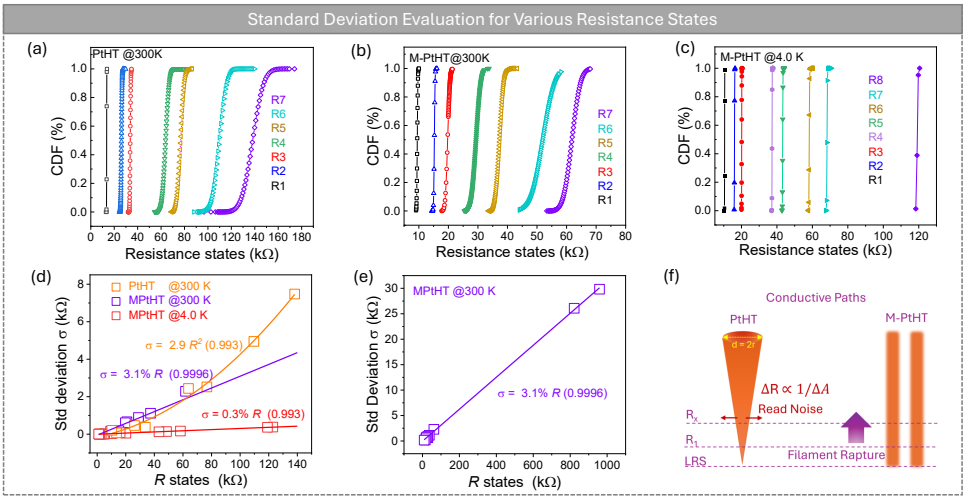
### 6.2.4. STATISTICAL VARIABILITY OF MULTI-BIT RESISTANCE STATES

Figure 6.6 quantifies the standard deviation of programmable resistance levels in PtHT and M-PtHT devices at 300 K and 4.0 K. Cumulative distribution functions (CDFs) of resistance levels are evaluated in Fig. 6.6(a) for PtHT at 300 K and it shows broad distributions for each level (20 kΩ–180 kΩ), with clear separation but substantial overlap at higher states. Meanwhile, M-PtHT at 300 K in Fig. 6.6(b) exhibits tighter clustering (10 kΩ–70 kΩ), indicating reduced read noise. Fig. 6.6(c) of M-PtHT at 4.0 K demonstrates the narrowest distributions, confirming cryogenic suppression of ionic motion and drift. Standard Deviation vs. Mean Resistance




Figuur 6.5: Endurance and multi-bit retention of PtHT and memristor-integrated PtHT (M-PtHT) devices at room temperature and cryogenic conditions. **(a)** Endurance test of PtHT devices at 300 K with read bias 0.2 V, showing stable low- and high-resistance states over $> 10^3$ cycles. **(b)** Seven-level retention of PtHT at 300 K: each resistance level R1–R7 remains distinct and drift-limited over 800 s. **(c)** M-PtHT at 300 K similarly retains seven programmed levels with minimal overlap. **(d)** M-PtHT at 4.0 K demonstrates robust multi-level retention under cryogenic operation, confirming suitability for quantum-dot gate biasing.

plotting $\sigma$ against mean $R$ for PtHT (orange), M-PtHT at 300 K (purple), and M-PtHT at 4.0 K (red) reveals distinct trends: PtHT follows an exponential scaling $\sigma = 2.9\,R^2$ (COD=0.993), reflecting increasing read noise at high resistance. M-PtHT at 300 K and 4.0 K follow linear relationships $\sigma = 0.031\,R$ (COD=0.9996) and $\sigma = 0.003\,R$ (COD=0.993), respectively, indicating voltage-programming noise remains proportional to the state value and is suppressed at cryogenic temperatures. Fig. 6.6(e) show expanded M-PtHT linearity for M-PtHT at 300 K, highlighting their sub-percent slopes (read noise rate). These results confirm that the M-PtHT stack achieves the lower-noise, multilevel stability, necessary for sub-100-µV bias generation in quantum-dot control circuits as simulated results in Fig. 6.2(d). In PtHT, read noise is dominated by stochastic trap capture, emission and local Joule-heating in nanometric filament constrictions. Because $R \propto 1/A$, small cross-section fluctuations $\Delta A$ induce $\Delta R \propto R^2$, yielding quadratic scaling. With a 1 nm Pt interlayer (M-PtHT), vacancy migration is homogenized across many conductive paths; numerous uncorrelated $\Delta R$ events aggregate Gaussian-like, so $\sigma$




Figuur 6.6: Statistical variability of multi-bit resistance states in PtHT and M-PtHT memristors. **(a–c)** CDFs of seven programmed levels R1–R7 at 300 K in PtHT (a), 300 K in memristor-integrated M-PtHT (b), and 4.0 K in M-PtHT (c), showing tighter clustering in the M-PtHT devices. **(d)** Standard deviation $\sigma$ versus mean resistance: PtHT at 300 K (orange) exhibits an exponential increase, fitted by $\sigma = 2.9\,R^2$, whereas M-PtHT at 300 K (purple) and 4.0 K (red) follow linear trends with slopes of 3.1% and 0.3%, respectively. **(e)** Expanded M-PtHT linearity and it fits for M-PtHT at 300 K. **(f)** Schematic illustration of the read noise difference for PtHT and M-PtHT.

scales linearly with $R$. Cryogenic cooling reduces trap kinetics, lowering the linear noise coefficient while preserving proportional dependence (Fig. 6.6(f)).

## 6.3. CONCLUSION OF THIS CHAPTER

We have demonstrated interface-enhanced Pt/$HfO_2$/Ti/Pt memristors that deliver eight distinct resistance states with approximately 0.3 % read noise rate. Minimal 6 memristors being integrated into a cryogenic PGA with $R_L = 30\,\text{k}\Omega$ and $V_{in} = 0.3\,\text{V}$, can guarantee a ±1 V range and <100 µV resolution per channel. This approach will collapse wiring complexity to a single bias feed and serial interface based on our initial simulation and single device performance evaluation, enabling scalable, high-precision control for next-generation quantum processors.

# 7

# MEMRISTOR-BASED NEURAL NETWORKS FOR QUANTUM STATE TOMOGRAPHY

*QUANTUM State Tomography (QST) is essential for characterizing and validating quantum systems, but its practical use is severely limited by the exponential growth of the Hilbert space and the number of measurements required for informational completeness. Many prior claims of performance have relied on architectural assumptions rather than systematic validation. We benchmark several neural network architectures to determine which scale effectively with qubit number and which fail to maintain high fidelity as system size increases. To address this, we perform a comprehensive benchmarking of diverse neural architectures across two quantum measurement strategies to evaluate their effectiveness in reconstructing both pure and mixed quantum states. Our results reveal that CNN and CGAN scale more robustly and achieve the highest fidelities while Spiking Variational Autoencoder (SVAE) demonstrates moderate fidelity performance, making them strong candidates for embedded, low-power hardware implementations. Recognizing that practical quantum diagnostics will require embedded, energy-efficient computation, we also discussed how memristor-based Computation-in-Memory (CiM) platforms can accelerate these models in hardware, mitigating memory bottlenecks and reducing energy consumption to enable scalable in-situ QST. This work identifies which architectures scale favorably for future quantum systems and lays the groundwork for quantum–classical co-design that is both computationally and physically scalable.*

The contents of this chapter were based on the preprint arXiv paper [149].

## 7.1. INTRODUCTION OF THIS CHAPTER

QUANTUM computing represents a groundbreaking paradigm that promises to redefine the boundaries of information processing. It leverages quantum superposition and entanglement to solve classically intractable problems in cryptography, simulation, and optimization [150–154]. Therefore, understanding the quantum system is necessary to form a critical operational foundation for calibration, error detection, benchmarking, and validation of quantum devices and algorithms with QST from measured data [155]. Quantum state tomography (QST) is the process of reconstructing the complete quantum state of a system based on a series of measurements [156, 157] . As quantum systems grow larger and increasingly complex, quantum state tomography (QST) becomes vital for verifying and benchmarking quantum devices [158]. In a world where quantum computers and neuromorphic computing systems are rapidly emerging, traditional QST methods, hampered by the exponential scaling of the Hilbert space, face significant computational challenges. For an $N$-qubit system, each qubit is associated with a two-dimensional Hilbert space, $\mathscr{H}_2$. The overall Hilbert space of the system is given by the tensor product:

$$\mathscr{H} = \mathscr{H}_2^{\otimes N},$$

which has dimension $2^N$. Consequently, any operator acting on this space, including the density matrix $\rho$ that describes the quantum state, must be a $2^N \times 2^N$ matrix. However, QST faces a central challenge: exponential scaling in both the Hilbert space ($2^N$ for $N$ qubits) and required measurement bases ($4^N$), which hampers its practicality for large-scale quantum systems [19, 157].

7

To overcome this bottleneck, a complementary path has emerged: leveraging machine learning to reduce measurement or computational overhead. Recent advances in applying artificial intelligence (AI), particularly neural networks [159], to QST have demonstrated strong potential for mitigating the curse of exponential dimensionality, by learning to reconstruct quantum states from fewer, noisy, or incomplete measurements [160, 161]. Numerous recent studies have demonstrated the effectiveness of various neural network architectures, including Convolutional Neural Networks (CNN) [51, 162, 163], Fully Connected Networks (FCN), Recurrent Neural Networks (RNN) [164], Restricted Boltzmann Machines (RBM) [165], Conditional Generative Adversarial Networks (CGAN) [166], Transformers [167, 168], and Variational Autoencoders (VAE) [169, 170]. While recent studies demonstrate that neural networks–based QST can be effective for small numbers of qubits, scaling these approaches to larger, practical quantum systems remains a challenge. This scalability demands not only algorithmic efficiency but also energy-efficient hardware support. However, current software-centric methods rarely address these hardware constraints. Addressing these critical limitations necessitates a shift toward hardware-aware neural network architectures. Conventional von Neumann computing architectures, characterized by separated memory and processing units, are severely limited by the *memory wall* problem, resulting from substantial data transfer bottlenecks that constrain computational efficiency and scalability [6]. CiM, particularly utilizing memristors technology, offers an innovative alternative. It

integrates memory storage and data processing capabilities within a single device, enabling improvements in energy efficiency, speed, and scalability by minimizing data movement and enabling analog computation [15, 171, 172].

In this work, we aim to identify which neural network architectures are scalable, accurate, and hardware-compatible for QST, particularly as quantum systems grow in size and complexity. To that end, we comprehensively benchmarked a diverse set of neural network architectures, supervised models (CNN, FCN, RNN, CGAN, Transformer) and unsupervised models (RBM, SVAE), to assess their suitability for reconstructing high-dimensional quantum states. Beyond simply comparing performance, our goal was to uncover which models maintain high fidelity, converge quickly, and scale favorably as the number of qubits and measurement complexity increase. Among them, CGAN and CNNs consistently outperform others, achieving fidelity up to 0.995 while offering fast convergence and computational efficiency. We report, additionally, the application of the Spiking Variational Autoencoder (SVAE) to QST. Unlike previous DNN-based models, SVAE leverages a sparse, event-driven architecture inspired by neuromorphic computing. Our results show that SVAE achieves high reconstruction fidelity while requiring significantly fewer computational resources. This makes it a strong candidate for future QST platforms, such as edge or embedded quantum diagnostic tools.

## 7.2. BACKGROUND

### 7.2.1. MEASUREMENT FORMALISM



In QST, measurement data are obtained by performing well-defined quantum measurements on an ensemble of identically prepared quantum states. Among the most widely used schemes, particularly in theoretical QST, are *projective measurements*, which correspond to projections of a quantum state onto the eigenbasis of a Hermitian operator. For single-qubit systems, the standard measurement basis is defined by the Pauli operator set:

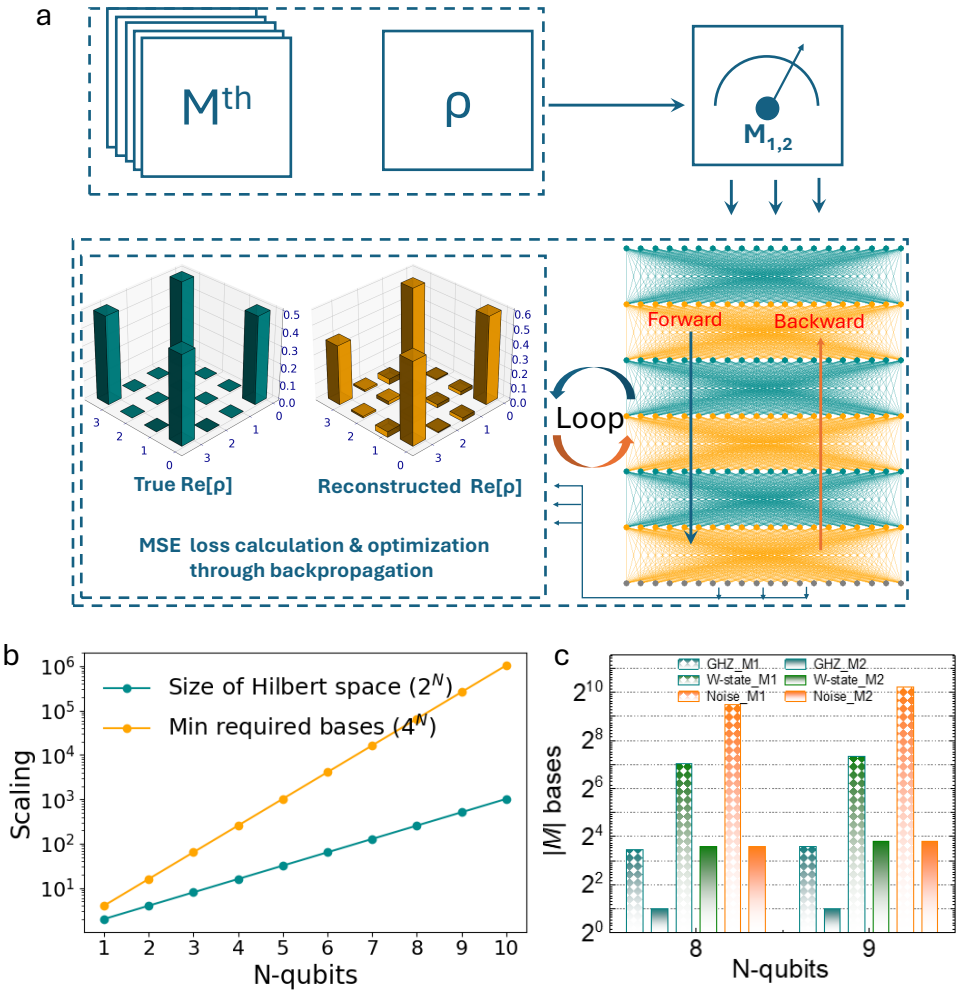
$$\{\sigma_x, \sigma_y, \sigma_z, \mathbb{I}_2\}, \tag{7.1}$$

where $\mathbb{I}_2$ is the $2 \times 2$ identity operator. These operators form a complete orthonormal basis for the space of Hermitian operators acting on $\mathbb{C}^2$. For $N$-qubit systems, the Pauli basis generalizes via tensor products of single-qubit operators, yielding $4^N$ distinct measurement operators. For instance, for two qubits, a representative element is:

$$\sigma_z \otimes \sigma_z. \tag{7.2}$$

Projective measurements are a special case of a more general framework known as *Positive Operator-Valued Measures* (POVMs). A POVM is defined by a collection of positive semi-definite operators $\{M_a\}$ satisfying the completeness relation:

$$\sum_a M_a = \mathbb{I}_d, \tag{7.3}$$




Figuur 7.1: **Overview of Neural Network-based QST.** (a) Neural Network-based workflow for reconstructing quantum states, optimizing through MSE loss calculation and backpropagation. (b) Scaling of Hilbert space size ($2^N$) and minimum required measurement bases ($4^N$) with increasing qubit number $N$. (c) Measurement bases required for high-fidelity ($\approx 0.99$) reconstruction of GHZ, W, and Noise states using expectation-value (method 1 or M1) and eigenstate-probability (method 2 or M2) methods.

where $\mathbb{I}_d$ is the identity operator on the Hilbert space of dimension $d = 2^N$. The probability of obtaining outcome $a$ when measuring the quantum state $\rho$ is given by:

$$P(a) = \mathrm{Tr}(\rho M_a). \tag{7.4}$$

To enable full quantum state reconstruction, the measurement operators must be *informationally complete* (*IC*), meaning that their statistical outcomes are sufficient to uniquely determine any $\rho$. A set of measurement operators $\{M_a\}$ is *IC* if it spans the space of linear operators on $\mathscr{H}_d$. That is, any operator $|\lambda\rangle$ in this space can be expressed as a linear combination of the measurement vectors:

$$|\lambda\rangle = a|\alpha\rangle + b|\beta\rangle + c|\gamma\rangle + \cdots . \tag{7.5}$$

In practice, due to finite sampling and noise, this reconstruction is achieved only approximately, and the accuracy depends on the number of measurements, qubit decoherence, and the reconstruction method used. In QST experiments, a large number of identically prepared quantum systems are measured under different settings to gather statistics. The two main types of information extracted from such measurements are: i) The *expectation value* of observables. ii) The *probability distribution* over measurement outcomes. For a pure quantum state $|\psi\rangle$, the expectation value of a Hermitian observable $\hat{A}$ is given by:

$$\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle. \tag{7.6}$$

For a mixed state described by a density matrix $\rho$, this generalizes to:

$$\langle A\rangle = \mathrm{Tr}(\rho\hat{A}). \tag{7.7}$$

These expressions return the average eigenvalue associated with the measurement of $\hat{A}$. For instance, computing $\langle\sigma_x\otimes\sigma_x\rangle$ reveals the expectation value for a two-qubit measurement in the $X$-basis.

In parallel, one can analyze the full probability distribution of outcomes. For a pure state $|\psi\rangle$, the probability of finding the system in eigenstate $|a\rangle$ is:

$$P(|a\rangle) = |\langle a|\psi\rangle|^2. \tag{7.8}$$

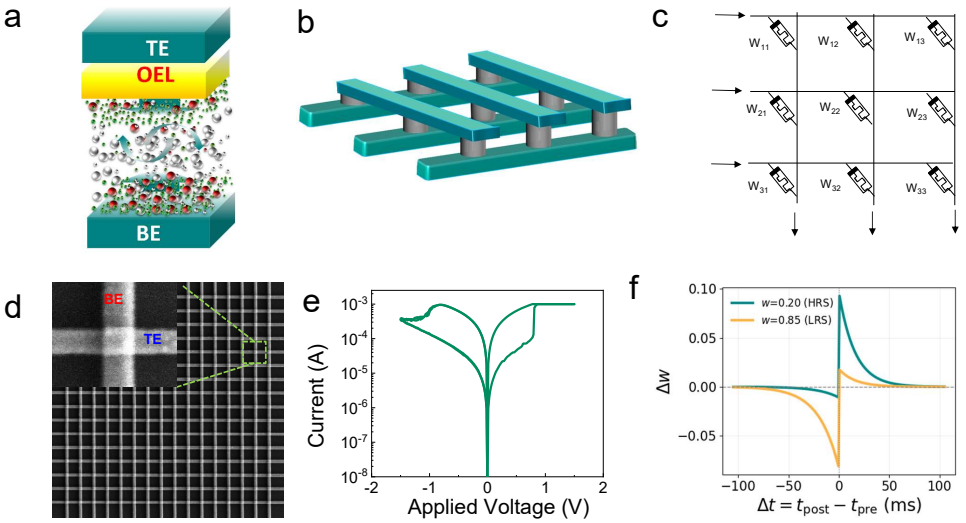
For mixed states, this probability becomes:

$$P(|a\rangle) = \mathrm{Tr}(\rho|a\rangle\langle a|). \tag{7.9}$$

These measurement statistics, expectation values or full probabilities, form the foundation of quantum state reconstruction. Whether via maximum likelihood estimation, Bayesian inference, or machine-learning–based techniques, all QST methods ultimately rely on the informational completeness of the chosen measurement protocol.

### 7.2.2. NEURAL NETWORK ARCHITECTURES AND LEARNING PARADIGMS

To address the challenge of reconstructing quantum states from measurement data, we explore artificial neural networks as efficient learning-based models. We consider seven representative neural networks architectures in this study: CNNs, FCNs, RNNs, RBMs, CGANs, Transformers and SVAE, chosen to span a broad spectrum of




Figuur 7.2: **a** Illustrative diagram of a memristor. TE: top electrodes; OEL: oxygen change layer; BE: bottom electrodes; the red, green and grey dots represent metals from BE or TE, oxgyen ions and oxygen vacancies, respecitvely. **b** Schematic crossbar array of memristor 3 × 3 and its circuitry reprsentation for MVM computation (**c**). **d** Scanning electron Microscopy (SEM) images of a real fabricated 16 × 16 memristor crossbar array. **e** I-V plot of a memristor. **f** STDP learning rules for synaptic plasticity

7

learning paradigms (supervised vs. unsupervised), structural designs (feedforward, recurrent, generative, spiking), and application strengths (e.g., spatial encoding, temporal modeling, distribution learning). For detailed architectural descriptions, see Appendix 7.5.2. Each model offers unique inductive biases tailored to specific learning tasks, for instance, CNNs for spatially structured inputs, RNNs for sequential data, and Transformers for attention-driven context modeling. While architectural design is important, the ultimate performance of a neural network is predominantly determined by the nature and quality of the training data. Equally crucial is the learning paradigm, such as supervised, unsupervised, or generative training, which is typically intrinsic to the architecture itself and significantly shapes its behavior. *Supervised Learning:* The most common paradigm, it utilizes labeled datasets to train a model by minimizing a predefined loss function that quantifies the difference between the predicted and actual outputs. *Unsupervised Learning:* This approach relies on unlabeled data, with the objective of discovering latent structure or statistical patterns, such as correlations, clusters, or low-dimensional manifolds, that characterize the data distribution [173].

In this work, we focus on supervised and unsupervised learning paradigms for quantum state reconstruction, as they are the most established and practically applicable frameworks in this domain. Reinforcement learning and other paradigms remain less explored in QST and are therefore beyond the scope of this study. In

the unsupervised setting, models learn a probability distribution from measurement data and subsequently reconstruct the corresponding quantum state. On contrast, supervised learning directly maps measurement data to a target quantum state, allowing direct mappings from measurement data to target quantum states, which enables task-specific training objectives and more data-efficient optimization by minimizing supervised loss functions. Among the architectures considered, all neural networks models employ supervised learning, with the exception of the *RBM* and the *SVAE* models, which are trained using unsupervised techniques as listed in the table 7.1.

Tabel 7.1: Overview of neural network architectures, learning paradigms, and DNN classification.

| **Model** | **Learning Paradigm** | **DNN-based** | **Notes** |
|---|---|---|---|
| CNN | Supervised | Yes | Spatial inductive bias |
| FCN | Supervised | Yes | Fully connected layers |
| RNN | Supervised | Yes | Temporal sequences |
| CGAN | Supervised | Yes | Generative supervised mapping |
| Transformer | Supervised | Yes | Attention mechanism |
| RBM | Unsupervised | No | Energy-based generative model |
| SVAE | Unsupervised | No | Spiking, event-driven encoding |

Training of neural networks proceeds through two fundamental phases: *feed-forward computation* and *backpropagation*. In the feed-forward phase, input data are propagated through the network layers to generate an output. In the backpropagation phase, the model prediction is compared against the ground truth using a loss function, and the resulting error gradient is propagated backward through the network to update the model parameters. This process is repeated iteratively until convergence, i.e., when the loss is minimized below a specified threshold [174]. The workflow is illustrated in Fig. 7.1(a), where it provides neural network–based approach to QST. To operationalize this, we outlines the neural network training process for QST in Fig 7.1(a). Firstly, theoretical measurement bases ($M^{th}$) and density matrix ($\rho$) are used to generate simulated measurement data ($M_{1,2}$). This data is fed into the neural network, which outputs the reconstructed state ($\rho$) as shown in 3D plots which illustrate true and reconstructed real parts (Re[$\rho$]) of density matrices for fidelity evaluation. Optimization of the network parameters is typically performed using stochastic gradient descent (*SGD*). In practice, we employ the *Adam optimizer*, a variant of *SGD* that uses first-order gradient estimates combined with adaptive learning rates and moment estimates for more efficient convergence. Adam requires relatively little memory and is widely used for deep learning applications [175].

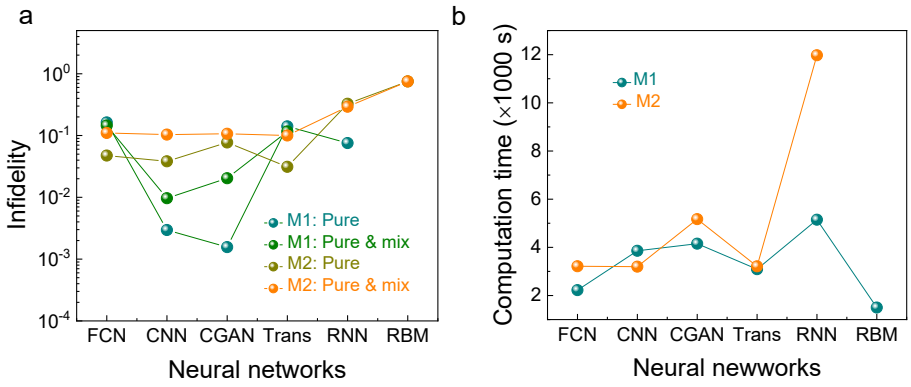
The choice of loss function is application-specific. For quantum state reconstruction, we adopt the commonly used *Mean Squared Error (MSE)* loss, which computes the average of the squared differences between predicted outputs $\hat{y}_i$ and

true values $y_i$ across a dataset of $N$ samples:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{7.10}$$

This loss is both simple to implement and analytically tractable, making it a natural choice for regression-type learning problems such as quantum state estimation.

The methodology used for reconstructing quantum states in this study involves training a neural network to map measurement data to a target quantum state. In the pre-processing stage, for example, a 3-qubit GHZ state from Eq. 7.14 and Pauli basis $XYZ$, $XIX$ and $ZXY$ are generated. Subsequently, measurements are performed on the GHZ state with the operators using Eq. 7.6,7.7 to compute the expectation value of measuring the three observables and Eq. 7.8, 7.9 to compute the probability of finding a quantum state in the eigenstate $|a\rangle$. In the training loop stage the resulting measurement data is used as input to a neural network, which after transformations performed by the hidden layers will output the complex coefficients of the reconstructed quantum state. Measurements are performed on the reconstructed state and used to compare to the true measurement data obtained during the pre-processing stage. Subsequently, the measurement outcomes are used to minimize the MSE loss function from Eq. 7.10 during each loop of the training stage.




Figuur 7.3: Infidelity **(a)** and computation time **(b)** over 100 iterations in reconstructing pure and mixed quantum states with two different methods to acquire measurement data for different neural network architectures.

### 7.2.3. MEMRISTORS-BASED ENERGY-EFFICIENT COMPUTING FOR SCALABLE QST

Computation in Memory (CiM) is a promising paradigm designed to overcome the memory wall problem associated with conventional von Neumann architectures. Traditional systems require frequent data transfers between memory and processor,

resulting in significant latency and energy inefficiency. In contrast, CiM architectures enable both storage and computation in the same physical location, thereby reducing data movement and improving computational throughput.

ReRAM, also known as a memristor, is a two-terminal resistive device that naturally aligns with the CiM paradigm. These devices can function as both memory and computational units, making them ideal candidates for energy-efficient, non-von Neumann architectures. Compared to CMOS technology, ReRAMs offer key advantages including non-volatility, low power consumption, small footprint, high scalability, and fast analog computation capabilities [176, 177]. Figure 7.2(a) shows the basic structure of a memristor, comprising a metal/insulator/metal stack where the insulating layer (typically an oxide) is sandwiched between top (TE) and bottom (BE) electrodes. Figures 7.2(b) and (c) illustrate a 3×3 memristor crossbar array designed to perform matrix-vector multiplication (MVM), a core operation in neural networks. Mathematically, MVM is given by:

$$\mathbf{y} = \mathbf{W} \cdot \mathbf{x} \tag{7.11}$$

In this architecture: (i) Each memory cell stores a weight ($\mathbf{W}$) as its conductance; (ii) The input vector $\mathbf{x}$ is applied as voltages across word lines. (iii) The resulting current at each bit line inherently performs analog multiply-and-accumulate (MAC) operations governed by Ohm's and Kirchhoff's laws. Figure 7.2(d) shows an SEM image of a fabricated 16×16 crossbar array with 100 nm × 100 nm node dimension, demonstrating physical feasibility. Compared to digital hardware, such analog computation offers key advantages for VMM-intensive applications like QST. Specifically, ReRAM-based CiM that is used to perform VMMs has the following potential: (i) *Reduced computation time:* Analog MAC operations replace sequential digital steps, allowing parallel execution of entire matrix-vector operations in a single cycle. (ii) *Lower energy consumption:* By eliminating the need for memory access and reducing data movement, ReRAM-based VMM consumes significantly less power per operation. (iii) *Massive parallelism:* All weights and inputs are operated on simultaneously in the crossbar, ideal for the parallel nature of quantum state reconstructions. (iv) *Improved scalability:* As the number of qubits increases, so does the model complexity. ReRAM small footprint and stackable architecture allow scaling to meet these growing demands.

In QST, neural networks are trained to reconstruct high-dimensional quantum states from measurement data. As qubit number $N$ increases, both the Hilbert space ($2^N$) and required measurement bases ($4^N$) grow exponentially. The efficiency of analog MVMs using CiM hardware thus becomes essential to sustain this scalability. ReRAM devices also exhibit binary resistance states: a low-resistance state (LRS, logic `1`) and a high-resistance state (HRS, logic `0`), as shown in Figure 7.2(e). Transitions between these states via SET and RESET operations underpin their functionality for storage and computation. Beyond inference, neuromorphic computing using ReRAM supports on-chip learning through spike-timing-dependent plasticity (STDP) [176, 177]. As illustrated in Figure 7.2(f), ReRAM synapses adjust conductance based on temporal patterns of neural activity. Applied to QST, this supports: (i) *Real-time adaptive learning:* STDP enables QST networks to be updated on-chip as new

quantum measurements are obtained. (ii) *Energy-efficient optimization:* Local weight adaptation removes the need for high-latency, high-power global updates. (iii) *Scalable deployment:* Embedded STDP learning within ReRAM makes it feasible to deploy self-improving QST systems as the quantum system scales. Thus, by leveraging both analog inference and local learning, ReRAM-based CiM and STDP mechanisms align tightly with the computational demands of QST. This synergy offers a robust and energy-efficient hardware substrate for building scalable QST engines.

## 7.3. RESULTS

### 7.3.1. MEASUREMENT BASES

The actual number of measurement bases $|M|$ needed for accurate QST depends not only on the choice of measurement strategy, such as M1 or M2, but also on the type of quantum state, the neural network architecture used for reconstruction, and the target fidelity (e.g., up to 99%). Since the optimal $|M|$ for a given NN and state is generally unknown, this section empirically evaluates how these factors influence measurement requirements.

To understand the challenge of curse of dimensionality, Fig 7.1(b) visualizes the scaling of the Hilbert space dimensionality ($2^N$) and the minimal required measurement bases ($4^N$) as functions of qubit number $N$. We also compare methods M1 and M2 to understand how the type of measurement data affects the number of required measurement bases $|M|$ for accurate QST. The purpose of this analysis is to determine which method scales more favorably and under what conditions. Fig 7.1(c) compares measurement bases required to achieve near-unity fidelity ($\approx 0.99$) reconstruction for three distinct quantum states (GHZ, W-state, and Noise) using aforementioned two methods. Bar heights indicate measurement bases count, highlighting the significant efficiency gain achieved using eigenstate probability-based measurement (M2). We investigate two representative strategies for generating measurement data: A pure quantum state with only real-valued amplitudes theoretically requires just one measurement basis to be fully reconstructed (up to a desired fidelity threshold, e.g., 99%). This is because such states contain amplitude but no phase information. However, when considering complex valued quantum states, phase information becomes relevant and, depending on the complexity of the state, a certain number of measurement bases are required to fully reconstruct it. In these cases, the number of required measurement bases grows, depending on the complexity of the state and the architecture used for learning. We use two types of methods to acquire measurement data because they reflect common practical approaches in QST research and offer a tradeoff between computational complexity and reconstruction performance: *M1:* Compute true expectation values $\hat{A}$ for the set of measurement bases $|M|$ using equation 7.6 for pure states $|\psi\rangle$ and equation 7.7 for mixed states $\rho$. *M2:* Compute true probabilities for measuring eigenstates $|a_i\rangle$ using the $M$ sets of measurement bases with equation 7.8 for pure states $|\psi\rangle$ and equation 7.9 for mixed states $\rho$.

Using measurement bases that result in an expectation value of zero creates

instability in the reconstruction process. This instability arises because the measured outcomes fluctuate equally between +1 and −1, which reduces the signal-to-noise ratio and makes the neural network estimation of the expectation value highly sensitive to small sampling fluctuations. Our numerical experiments (see Fig. 7.1(c)) show that including these zero-expectation bases does not improve the reconstruction fidelity for method M1. In fact, M1 is particularly affected because it relies on a single scalar expectation value per measurement basis; a zero value carries minimal information about the state. By contrast, non-zero measurement bases contain more information about the state. When the expectation value is close to ±1, the measurement outcome is nearly deterministic, which imposes stronger constraints on the possible quantum states. For this reason, we restrict our evaluation to non-zero measurement bases, as they provide the most reliable measurement data and the highest information content. However, as shown in Fig. 7.1(b), QST is fundamentally limited by the exponential scaling of the Hilbert space ($2^N$) and the minimum required number of measurement bases ($4^N$) to be informationally complete [48]. Fig. 7.1(c) then evaluates how many non-zero measurement bases are required to fully reconstruct three representative pure quantum states using methods M1 and M2. This allows us to explicitly connect the informational content of the measurement bases with the empirical reconstruction requirements.

To place the measurement requirements into context, we now evaluate how the number of measurement bases $|M|$ scales for representative quantum states. Our goal is to assess whether sub-exponential scaling in $|M|$ and the associated measurement data can be achieved in practice, compared to the exponential upper bound of $4^N$ required for informational completeness. For $N = 8$ qubits, the maximum number of measurement bases in the Pauli basis is $4^8 = 65,536$, and for $N = 9$ it is $4^9 = 262,144$. Figure 7.1(c) summarizes the empirically determined values of $|M|$ required to reach a reconstruction fidelity of approximately 0.99 for three representative quantum states: a GHZ state, a W state, and a noisy pure state (see Appendix 7.5.1 for definitions). These $|M|$ values are obtained from our numerical experiments. For the GHZ state, method M1 requires $|M| = 11$ for $N = 8$ and $|M| = 12$ for $N = 9$, whereas M2 only requires $|M| = 2$ for both $N = 8$ and $N = 9$. For the W state, method M1 requires $|M| = 130$ for $N = 8$ and $|M| = 160$ for $N = 9$, while method M2 requires $|M| = 12$ and $|M| = 14$, respectively. For the noisy state, method M1 requires $|M| = 720$ for $N = 8$ and $|M| = 1,200$ for $N = 9$, whereas method M2 requires $|M| = 40$ for $N = 8$ (reconstruction for $N = 9$ was not feasible within available memory). We observe that as the number of computational basis states with non-zero amplitudes in the quantum state increases, the number of measurement bases required to fully reconstruct both the amplitudes and phases also increases. Interestingly, the degree of entanglement itself is not the dominant factor: although the GHZ state is maximally entangled, it is relatively easy to reconstruct, while the W state, despite being less entangled, requires significantly more measurement bases due to the larger number of computational basis states with non-zero amplitudes.

In summary, M2 generally requires fewer measurement bases with non-vanishing expectation values compared to M1. This is because the measurement data obtained in M2 are full probability distributions over the Hilbert space for each basis, whereas

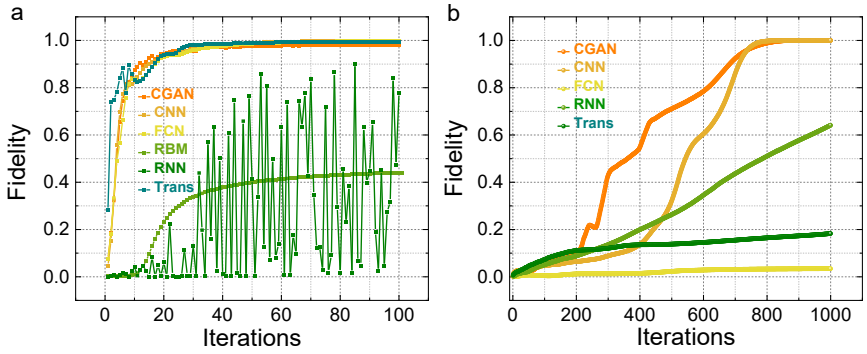
M1 only provides a single scalar expectation value per basis. As a result, each data set from M2 constrains the possible quantum states more strongly, which in turn reduces the number of distinct measurement bases needed for a given target fidelity. However, while this advantage is clear in simulation (where ideal probability distributions can be directly computed), it is less practical experimentally. In real experiments, obtaining the full probability distribution for M2 requires significantly more repeated measurements per basis to collect sufficient statistics for each possible outcome. In contrast, M1 only requires repeated measurements to estimate the average expectation value, which typically converges with fewer samples. This makes M1 more scalable in experimental settings, even though it requires more distinct measurement bases to reach the same reconstruction fidelity.

### 7.3.2. NEURAL NETWORK PERFORMANCE EVALUATION

To quantitatively assess the effectiveness of different neural network architectures for QST, we evaluate their performance across two critical metrics: reconstruction accuracy (fidelity) and computational efficiency. To identify which architectures maintain high fidelity while remaining computationally practical as quantum systems scale, and how different measurement strategies, expectation value-based ( M1) and eigenstate-probability-based (M2), affect reconstruction outcomes.

Figure 7.3 presents the average reconstruction infidelity and corresponding computation time across various neural network models, including FCN,CNN, CGAN, Trans, RNN, and RBM, for both pure and mixed quantum states under the two measurement methods with 8 qubits. As depicted in Fig. 7.3(a), two of six architectures CNN and CGAN consistently achieve the minimal infidelity values, with CGAN yielding the highest fidelity across all settings. Specifically, CGAN and CNN with M1 on pure states achieves the best reconstruction performance with a infidelity lower than $2 \times 10^{-3}$. In contrast, RBM and RNN perform poorly, especially for mixed states using M2, with infidelity values exceeding $10^{-1}$. While M2 is theoretically richer in information content, it leads to slightly higher reconstruction infidelity across most architectures, likely due to increased input complexity and learning instability when processing full probability distributions. In terms of computation Time as shown in Fig. 7.3(b), FCN and CNN demonstrate the shortest training durations (around 2,000 seconds), making them ideal for practical applications. CGAN and Transformer exhibit moderate computational demands (3,000–6,000 seconds), while RNN and RBM are significantly slower, particularly RBM under M2, which exceeds 12,000 seconds. This reinforces the need to consider both accuracy and efficiency when selecting architectures for real-world QST deployments. Taking above considerations, these results demonstrate that CNN offers the most balanced performance in terms of fidelity and speed. Although M2 theoretically provides more informative measurement data, M1 results in more stable and efficient training across most models, especially for mixed quantum states. This insight is crucial for guiding experimental and hardware-constrained implementations of neural-network-based QST.

To illustrate the details of reconstruction process for various neural networks. Figure 7.4(a) shows a 9 qubit GHZ state for different neural network architectures,




Figuur 7.4: Fidelity as a function of iterations for a noisy mixed state. **a** Reconstruction of a pure 9 qubit GHZ state for 6 different neural network architectures, using 2 non-zero measurement bases with M2. **b** Reconstruction process of a pure 9 qubit noisy state for 5 different neural network architectures, using 1200 non-zero measurement bases with M1.

Tabel 7.2: Computation time comparison (in seconds) for different neural network architectures using methods M1 and M2.

| **Neural Network** | **M2 (2 bases)** | **M1 (1200 bases)** |
|---|---|---|
| CGAN | 710 | 19,468 |
| CNN | 290 | 19,687 |
| FCN | 280 | 11,043 |
| RBM | 15,768 | N.A. |
| RNN | 843 | 21,252 |
| Trans | 303 | 16,798 |

performed with 2 non-zero measurement bases using M2. All neural networks rapidly converge to a fully reconstructed except for RNN and RBM. RNN shows major instability in the large oscillations between fidelity values and does not converge to a high fidelity ($> 0.99$). RBM only converges to a fidelity of around 0.43. For this case, supervised learning is able to perform significantly better than the unsupervised RBM model. However, while the supervised learning models are Deep Neural Networks (DNNs) with multiple (3 or more) hidden layers, the RBM model is only a single layer. This requires more research to reliably compare supervised and unsupervised learning models for QST. Among the neural networks, FCN has the smallest amount of interations and time for high-fidelity convergence, with CNN and Transformer showing very similar performance as listed in the caption. CGAN requires 2.5 times longer, RNN takes 3 times as long, and RBM is the most time consuming, taking 5.6 times. In Figure 7.4(b), a 9-qubit state initially prepared as a pure state is reconstructed after being subjected to noise, resulting in a mixed state. The reconstruction is performed with $|M| = 1,200$ measurement bases using

M1. Among the tested models, only CNN and CGAN are able to fully reconstruct the mixed state (fidelity $> 0.99$) within roughly the same number of iterations. In contrast, RNN, Transformer, and FCN do not reach comparable fidelities within 1000 iterations; FCN, in particular, almost fails to extract meaningful information from the measurement data. Notice how many more iterations are needed here compared to the noiseless case in Figure 7.4(a); this is due to the increased complexity of the noisy mixed state, which requires more measurement bases and longer training to reconstruct. CNN and CGAN have approximately equal computation times, as listed in Table 7.2.

These results highlight CNN and CGAN as the most robust and scalable supervised architectures, while unsupervised models like RBM struggle to achieve comparable fidelity. These insights provide a practical guideline for selecting models when balancing fidelity, computational time, and scalability in QST.

Because the SVAE uses an event-driven spiking architecture, it can be naturally mapped to neuromorphic hardware such as BrainScaleS-2 and ReRAM-based CiM accelerators. This makes it more hardware-friendly compared to standard CNNs or Transformers. Meanwhile, SVAE differs fundamentally from deterministic supervised models in both learning paradigm and hardware relevance, and shows unique fidelity-scaling trends, we evaluate it separately to highlight its strengths and limitations for scalable QST. Figure 7.5 presents an in-depth performance analysis of the *SVAE* architecture applied to QST using pure GHZ states with M2. The model performance is evaluated across two principal dimensions: the fidelity of reconstructed quantum states and the computational time required, both as a function of the number of qubits and the total number of measurement shots. Figure 7.5(a) shows the fidelity of the SVAE-generated quantum state reconstructions as a function of qubit number, evaluated for six total shot counts ranging from $10^1$ to $10^6$.

At low shot counts ($10^1$–$10^2$), the SVAE exhibits poor performance above 4 qubits, with fidelities rapidly decaying to near-zero values. This behavior is attributable to insufficient statistical sampling in high-dimensional Hilbert spaces, where the state space grows exponentially as $2^N$ for an $N$-qubit system. Without enough measurement data, the SVAE lacks the information required to learn a faithful generative distribution, resulting in high reconstruction error. As the number of measurement shots increases, the model performance improves markedly. For $10^5$ and $10^6$ shots, the SVAE consistently achieves fidelities above 0.9 for systems up to 6 qubits. These results highlight the SVAE capacity to utilize rich statistical information effectively. The saturation behavior observed in fidelity for high shot counts suggests that the model error becomes bounded by its expressivity rather than data limitations. Figure 7.5b displays the computational time required for the SVAE to perform quantum state reconstructions, plotted as a function of qubit number and shot count. For small systems (3 to 5 qubits), the inference time remains low, even at the highest shot levels. This computational efficiency stems from the SVAE ability to perform amortized inference, whereby the decoder maps latent representations to full density matrices without requiring iterative optimization for each individual measurement.

As both the number of qubits and the total shot count increase, the computational cost grows progressively. Larger datasets necessitate longer data loading and preprocessing times, and deeper networks or higher-dimensional latent spaces require more iterations during training and evaluation. For 7 to 8 qubits with $10^6$ shots, the time cost reaches the hundreds of seconds range, reflecting increased optimization complexity and the growing burden of processing high-dimensional input features. Unlike deterministic neural networks such as CNN, CGAN, and Transformer, the SVAE leverages its generative latent space to effectively model uncertainty and incomplete measurement data. Its probabilistic framework provides greater robustness under noisy or data-sparse conditions, where deterministic models often struggle. Moreover, the event-driven spiking architecture of SVAE makes it inherently more energy-efficient and hardware-friendly for neuromorphic and ReRAM-based CiM implementations, offering a scalable pathway for on-chip QST in resource-constrained environments.

Compared to traditional QST methods such as Maximum Likelihood Estimation (MLE) or Bayesian techniques, which scale exponentially with qubit number and become impractical beyond roughly 6–8 qubits [178, 179], the SVAE demonstrates a more favorable scaling profile. In our experiments, SVAE reliably reconstructed pure GHZ states up to 7 qubits using $10^5$–$10^6$ shots, whereas MLE would require a fully informationally complete measurement set scaling as $4^N$. Although fidelity drops at very low shot counts ($10^1$–$10^2$) for systems above 4 qubits (as shown in Fig. 7.5), SVAE generative modeling enables it to manage higher-dimensional states with fewer measurements and less computational overhead [169]. This balance positions SVAE as a promising candidate for near-term quantum experiments that require real-time feedback and practical reconstruction fidelity.



Figure 7.5 a,b illustrate a key trade-off in QST using SVAE models: fidelity improves with larger data availability, but at the cost of higher computation time. At high shot counts ($10^5$–$10^6$), the SVAE achieves the reasonable fidelities observed in our experiments; however, fidelity consistently declines as the number of qubits increases, and does not exceed 0.9 for systems above 4–5 qubits. At lower shot counts ($10^1$–$10^2$), the fidelity rapidly decreases for all qubit numbers, reflecting the severe information deficit when measurement data are sparse. These results underline that, while SVAE benefits from richer data, scalability remains a major challenge for systems with many qubits. One distinctive advantage of the SVAE, compared to deterministic neural network models such as CNNs, CGANs, or Transformers, is its spiking, event-driven architecture. This makes the SVAE inherently more hardware-friendly and energy-efficient, as it can be mapped onto neuromorphic platforms such as memristors-based computation-in-memory accelerators. Such compatibility positions SVAE as a promising candidate for energy-efficient, on-chip quantum state reconstruction in near-term experiments. These results reinforce broader conclusions in the literature that deep generative models with variational inference represent a compelling direction for mitigating the scalability barriers of conventional QST, while offering opportunities for more hardware-efficient implementations.

## 7.4. DISCUSSION OF THIS CHAPTER

This work presented a systematic benchmarking of neural network (NN) architectures for QST using two distinct measurement methodologies (M1 and M2). The comprehensiveness of the analysis stems from the inclusion of models across the major NN paradigms: FCN, CNN, RNN, transformers, CGAN, RBM, and SVAE. These models span supervised and unsupervised learning, deterministic and generative inference, and architectures with varying scalability and hardware compatibility, ensuring that the comparisons reflect the state of the art in machine learning for QST.

Several key observations emerge from this analysis. CNN architectures consistently achieve the best balance of reconstruction fidelity and computational efficiency across pure and mixed states, confirming their suitability for large-scale QST. CGAN also attains high fidelity but at higher computational cost. In contrast, RBM and RNN models are highly sensitive to hyperparameters (e.g., learning rate) and exhibit poor scalability, often failing to converge for larger qubit systems or mixed states. SVAE, a generative unsupervised model, shows distinctive behavior: at high shot counts ($10^5$–$10^6$), it attains competitive fidelities for small systems but fidelity declines steadily with increasing qubit number. Nonetheless, SVAE demonstrates enhanced robustness under noisy and data-sparse conditions, where deterministic supervised models degrade more severely. Furthermore, its spiking, event-driven architecture makes it inherently compatible with neuromorphic and ReRAM-based computation-in-memory (CiM) accelerators—an advantage not shared by CNNs or transformers.

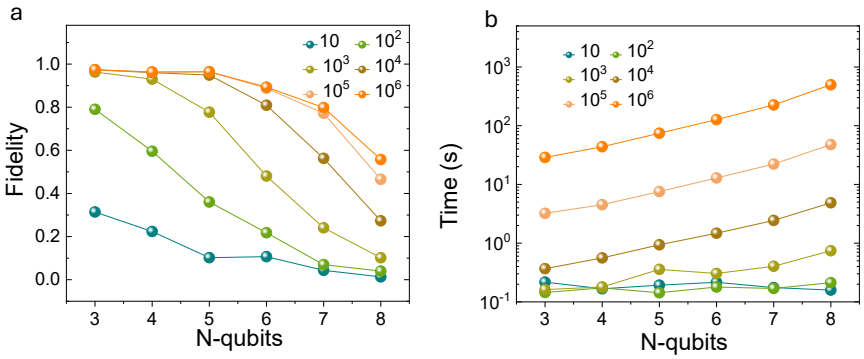
The comparison of measurement methodologies revealed that while M2 can, in principle, reduce the number of required measurement bases for pure states by exploiting informational completeness, M1 remains more practical for experimental mixed-state scenarios due to its lower measurement overhead per basis. For instance, in reconstructing a noisy $N = 6$ qubit mixed state, M1 achieved full reconstruction using the maximum $4^6 = 4,096$ bases, while M2 was constrained to only 80 bases for the same system. These findings highlight the interplay between measurement design and algorithmic scalability.

Beyond the algorithmic comparisons, we propose that future work should explore the mapping of CNN and SVAE architectures onto ReRAMs-based computation-in-memory (CiM) platforms. Such hardware-aware integration could, in principle, alleviate the computational bottlenecks of von Neumann architectures by reducing data movement, leveraging non-volatility, and exploiting analog computation. While this study did not include hardware-level simulations or implementations, the unique architectural properties of SVAE and CNN suggest that they are strong candidates for co-design with CiM accelerators to enable scalable, energy-efficient QST pipelines.

Looking forward, future work should move beyond incremental parameter tuning. One promising direction is the co-design of neural architectures and CiM hardware, jointly optimizing sparsity, quantization, and analog precision to further close the fidelity gap for large qubit numbers. Another is the development of hybrid supervised–unsupervised generative models capable of adaptively fusing simulated and experimental measurement data, thus improving robustness in realistic noise

environments. Finally, integrating these QST methods into real-time feedback loops could transform tomography from an offline diagnostic tool into an active component of quantum system calibration and error mitigation.

Taking all above considerations, this study establishes CNN as the most robust supervised architecture for QST, SVAE as a promising generative alternative with unique hardware compatibility, and M1 measurement strategies as the most practical for mixed-state reconstructions. By leveraging these insights and pursuing architecture–hardware co-design, the community can develop QST solutions that scale gracefully with system size, accelerating the transition to large-scale quantum information processing.




Figuur 7.5: **Performance evaluation of SVAE neural network for QST. a** Fidelity of the reconstructed quantum states as a function of the number of qubits (ranging from 3 to 8), evaluated over six different total measurement shot counts: $10^1$, $10^2$, $10^3$, $10^4$, $10^5$, and $10^6$. **b** Corresponding computational time required for the neural network training as a function of qubit number, under the same shot count conditions.

An important observation from our analysis was the notable sensitivity of certain neural network architectures, particularly the RBM, to variations in learning rate. This highlights the need for dedicated hyperparameter optimization to fully exploit the potential of unsupervised models. In contrast, supervised architectures such as CNN, FCN, and Transformer were more robust to learning rate variations, underscoring their practicality for large-scale QST.

To address computational limitations inherent in traditional von Neumann architectures, we conducted numerical experiments exploring the mapping of CNN and SVAE architectures onto ReRAM-based computation-in-memory (CiM) systems. These simulations demonstrate the theoretical performance gains possible when leveraging the non-volatility, analog computation, and reduced data movement of ReRAM devices. This hardware-aware perspective, combined with our neural network benchmarks, provides a pathway towards more scalable, energy-efficient QST.

Looking forward, future work should push beyond incremental hyperparameter tuning or standard loss function adjustments. One promising direction is to co-design neural architectures and CiM hardware for quantum applications, jointly

optimizing sparsity, quantization, and analog precision to further close the fidelity gap for large qubit numbers. In parallel, integrating generative and hybrid models that can adaptively fuse simulated and experimental measurement data may unlock new regimes of performance. Finally, exploring real-time deployment scenarios with error-mitigated hardware accelerators could transform QST from an offline diagnostic tool into a dynamic component of quantum system control.

Collectively, this study lays a foundation for such advances, establishing CNN as the most reliable supervised architecture, SVAE as a robust generative alternative with hardware compatibility, and M1 measurement strategies as the most practical for mixed states in experimental contexts. By building on these insights, the community can progress toward QST solutions that scale gracefully with system size, opening the door to more powerful quantum information processing.

## 7.5. METHODS OF THIS CHAPTER

The software used is written in Python version 3.11. The hardware specifications of the used computer are: 16 GB RAM, Intel i5-4460 CPU and GeForce GTX 1660 Super GPU with 6 GB of VRAM. A custom layer in the neural network model is used to extract the reconstructed quantum state in order to get a good measure of how the neural network model is performed by computing the fidelity of the true and reconstructed state. The fidelity, which is a measure used to compute the overlap between two quantum states is commonly used to indicate similarity between the states. The definition we used to calculate the fidelity of a pure state is:



$$Fidelity = |\langle\psi_1|\psi_2\rangle|^2. \tag{7.12}$$

The definition we used to calculate the fidelity of a mixed state is:

$$Fidelity = \left(\mathrm{Tr}\left[\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}}\right]\right)^2. \tag{7.13}$$

### 7.5.1. QUANTUM STATES

In general a quantum state consists of complex coefficients which contain two types of information: amplitude and phase information. The physical pure quantum states $\rho$ (with $\rho^2 = \rho$) of interest are the Greenberger-Horne-Zeilinger (GHZ) state, Wolfgang (W) state and a noise state. The GHZ state is maximally entangled for all number of $N$ qubits with a purity of 1. The GHZ state is defined as:

$$|GHZ\rangle = \frac{|0\rangle^{\otimes N} + |1\rangle^{\otimes N}}{\sqrt{2}}. \tag{7.14}$$

The W state contains a superposition of N qubits in which only one qubit in every ket is in the $|1\rangle$ state and the amount of entanglement decreases with increasing amount of qubits $N$. This state is maximally entangled for 2 qubits, partially entangled for 3 qubits and as the number of qubits $N$ increases the degree of

entanglement decreases. The W state is defined as:

$$|W\rangle = \frac{|100...0\rangle + |010...0\rangle + ... + |00...01\rangle}{\sqrt{N}}. \tag{7.15}$$

The pure noise state consists of randomly generated complex coefficients for every ket and is defined as:

$$|\psi\rangle = \sum_{i=0}^{2^N-1} c_i |i\rangle. \tag{7.16}$$

In order for the quantum states to be physical states, the corresponding density matrix requires to be positive semi-definite (PSD), hermitian and have $Tr(\rho) = 1$ (normalization).

The physical mixed states of interest are the generalized Werner (W) state, the mixed W state and a mixed noise state. The mixed W state will keep the same properties as the pure W state, but is now in the form of a 2D matrix of complex coefficients instead of a 1D vector. The mixed W state is defined as:

$$\rho = |W\rangle\langle W|. \tag{7.17}$$

The chosen generalized Werner state is defined as:

$$\rho = p|GHZ\rangle\langle GHZ| + (1-p)I_N/2^N, \tag{7.18}$$

which is a combination of the outer product of a GHZ state and a maximally mixed state coming from the second term $I_N/2^N$. For $p = 0$ it would purely be a maximally mixed state and for $p = 1$ it would purely be a mixed GHZ state. A maximally mixed state contains no amount of entanglement and gives a lower bound on the purity which decreases with increasing number of qubits as $1/2^N$. The overall purity of the Werner state with $p = 0.5$ decreases from 0.438 for 2 qubits to about 0.261 for 6 qubits and is partially entangled.

The mixed noise state also consists of randomly generated complex coefficients for every ket and is defined as:

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|, \tag{7.19}$$

where the density matrix $\rho$ again requires to be PSD, hermitian and have $Tr(\rho) = 1$. Comparing the results from reconstructing these three pure and three mixed quantum states will show the influence of the amount of computational basis states with non-zero amplitudes and the degree of entanglement on the performance of a neural network.

### 7.5.2. NEURAL NETWORKS ARCHITECTURES APPLIED IN THIS CHAPTER

#### FCN MODEL

Table 7.3 shows the specific layers used in the FCN model and the corresponding output shapes and trainable parameters in the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. The first input layer contains the true

Tabel 7.3: FCN model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer | (None, 4096) | 0 |
| InputLayer | (None, 64, 64, 8192) | 0 |
| Dense | (None, 2048) | 8,388,608 |
| Dense | (None, 2048) | 4,196,352 |
| Dense | (None, 4096) | 8,392,704 |
| Dense | (None, 4096) | 16,781,312 |
| Dense | (None, 8192) | 33,562,624 |
| Reshape | (None, 64, 64, 2) | 0 |
| DensityMatrix | (None, 64, 64) | 0 |
| Expectation | (None, 4096) | 0 |

expectation values for the $|M| = 4,096$ sets of measurement bases. The second input layer contains the measurement operators which are only used in the last custom layer Expectation. The shape value $8,192 = 4,096 \cdot 2$ is due to the separation of real and imaginary parts. Then five consecutive Dense layers are used in order to extract features from the input data and the Dense layers increase in complexity and hence will learn increasingly complex patterns. After each Dense layer a LeakyReLu activation is used to introduce non-linearity into the model which allows the model to learn about more complex features. The learned features are then used to make predictions. The last Dense layer is used for transforming the data into the desired output shape of $2 \cdot 64^2 = 8,192$ which is required for constructing the density matrix and hence cannot be reduced in size. First, the data needs to be reshaped into the proper form such that it can be passed to the custom DensityMatrix layer in which the reconstructed density matrix is computed. Finally, the expectation layer computes the new expectation values of the reconstructed density matrix.



### CNN MODEL

Table 7.4 shows the specific layers used in the CNN model and the corresponding output shapes and trainable parameters for the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. First, the data passes through a Dense layer to increase dimensionality, followed by a LeakyReLU activation to introduce non-linearity into the learning process. The data is then reshaped into a 4D format suitable for the following convolutional layers. The Conv2DTranspose layers are used for upscaling the dimensionality of the data and extracting data features, instead of the standard downscaling approach used in typical convolutional layers. This upscaling strategy is commonly employed for data generation or reconstruction tasks. The InstanceNormalization layers are applied after each transpose convolution to keep the data stable and normalized, compensating for the multiple up- and down-scaling steps in the architecture. Finally, the custom `DensityMatrix` and

Tabel 7.4: CNN model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer | (None, 4096) | 0 |
| InputLayer | (None, 64, 64, 8192) | 0 |
| Dense | (None, 2048) | 8,388,608 |
| LeakyReLU | (None, 2048) | 0 |
| Reshape | (None, 32, 32, 2) | 0 |
| Conv2DTranspose | (None, 64, 64, 64) | 2,048 |
| InstanceNormalization | (None, 64, 64, 64) | 128 |
| LeakyReLU | (None, 64, 64, 64) | 0 |
| Conv2DTranspose | (None, 64, 64, 64) | 65,536 |
| InstanceNormalization | (None, 64, 64, 64) | 128 |
| LeakyReLU | (None, 64, 64, 64) | 0 |
| Conv2DTranspose | (None, 64, 64, 2048) | 32,768 |
| Conv2DTranspose | (None, 64, 64, 2) | 1,024 |
| DensityMatrix | (None, 64, 64) | 0 |
| Expectation | (None, 4096) | 0 |

`Expectation` layers are appended to generate the final quantum state representation and measurement expectations.

### RNN MODEL



Tabel 7.5: RNN model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer (inputs) | (None, 4096) | 0 |
| InputLayer (operators) | (None, 64, 64, 8192) | 0 |
| Reshape | (None, 4096, 1) | 0 |
| SimpleRNN | (None, 4096, 50) | 2,600 |
| SimpleRNN | (None, 50) | 5,050 |
| Dense | (None, 8192) | 417,792 |
| Reshape | (None, 64, 64, 2) | 0 |
| DensityMatrix | (None, 64, 64) | 0 |
| Expectation | (None, 4096) | 0 |

Table 7.5 shows the specific layers used in the RNN model and the corresponding output shapes and trainable parameters in the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. The difference between the FCN model and this one is that the Dense layers are replaced by two SimpleRNN layers. The

first SimpleRNN layer processes the input sequentially and feeds the output in a sequence to the second SimpleRNN layer. The second layer processes the sequence of data and returns the final output of the two layers. Here, the Tanh activation function is used, which is more commonly applied in RNN models to introduce non-linearity such that the model is again able to learn more complex patterns.

### RBM MODEL

Tabel 7.6: RBM model summary

| **Model detail** | **Value** |
|---|---|
| Number of visible units | 6 |
| Number of hidden units | 13 |
| Number of parameters in $\lambda$ weights | 162 |
| Number of parameters in $\mu$ weights | 162 |
| Total parameters | 324 |

This model uses unsupervised learning and is not part of the DNN category because it only has two layers: a visible and a hidden layer. The RBM model is used to learn a probability distribution from measurement data and subsequently uses the probabilities to reconstruct the quantum state it represents. The parameters used in training the RBM model are shown in Table 7.6. The RBM model generates counts for every configuration, which are subsequently converted into probabilities. Unlike the M2 measurement methodology used in other neural network models, where exact eigenstate probabilities are provided without sampling noise, the RBM uses measurement data that inherently includes statistical fluctuations. The $\lambda$ weights are used for the amplitudes and the $\mu$ weights for the phases of the quantum state, and both sets of parameters are updated in every loop by computing gradients. The number of hidden neurons is set at $2N+1$, and a default of 1,000 counts are generated per basis. The initial learning rate is set high at 0.1 to prevent the model from being stuck at zero fidelity during early training.

### CGAN MODEL

CGAN consists of a generator and a discriminator model. Table 7.7 shows the specific layers used in the Generator model and the corresponding output shapes and trainable parameters in the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. This model is equivalent to the CNN model.

Table 7.8 shows the specific layers used in the Discriminator model and the corresponding output shapes and trainable parameters in the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. This model does not differ significantly from the FCN model, only that another input is present which is the input image generated with the Generator model. Also, a Concatenate layer is used to merge the two input layers into a single tensor.

Tabel 7.7: CGAN: Generator model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer | (None, 4096) | 0 |
| InputLayer | (None, 64, 64, 8192) | 0 |
| Dense | (None, 2048) | 8,388,608 |
| LeakyReLU | (None, 2048) | 0 |
| Reshape | (None, 32, 32, 2) | 0 |
| Conv2DTranspose | (None, 64, 64, 64) | 2,048 |
| InstanceNormalization | (None, 64, 64, 64) | 128 |
| LeakyReLU | (None, 64, 64, 64) | 0 |
| Conv2DTranspose | (None, 64, 64, 64) | 65,536 |
| InstanceNormalization | (None, 64, 64, 64) | 128 |
| LeakyReLU | (None, 64, 64, 64) | 0 |
| Conv2DTranspose | (None, 64, 64, 32) | 32,768 |
| Conv2DTranspose | (None, 64, 64, 2) | 1,024 |
| DensityMatrix | (None, 64, 64) | 0 |
| Expectation | (None, 4096) | 0 |

Tabel 7.8: CGAN: Discriminator model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer (input image) | (None, 4096) | 0 |
| InputLayer (target image) | (None, 4096) | 0 |
| InputLayer (operators) | (None, 64, 64, 8192) | 0 |
| Concatenate | (None, 8192) | 0 |
| Dense | (None, 128) | 1,048,704 |
| LeakyReLU | (None, 128) | 0 |
| Dense | (None, 128) | 16,512 |
| LeakyReLU | (None, 128) | 0 |
| Dense | (None, 64) | 8,256 |
| Dense | (None, 64) | 4,160 |



### TRANSFORMER MODEL

Table 7.9 shows the specific layers used in the Transformer model and the corresponding output shapes and trainable parameters in the case of $N = 6$ qubits and $|M| = 4^6 = 4,096$ measurement bases. The deviating layer compared to the previous models is the TransformerEncoder layer. Essentially, the encoder processes sequential data using self-attention to focus on the most significant parts of the data, potentially capturing features more efficiently. Since the decoder from the standard Transformer architecture is not required for QST, it is omitted, simplifying

the model and improving efficiency. We set the number of attention heads to 8 and the number of encoder layers to 4.

Tabel 7.9: Transformer model summary

| Layer type | Output shape | Parameters |
|---|---|---|
| InputLayer | (None, 4096) | 0 |
| InputLayer | (None, 64, 64, 8192) | 0 |
| TransformerEncoder | (1, 128, 128) | 856,960 |
| Flatten | (1, 16,384) | 0 |
| Dense | (1, 8192) | 134,225,920 |
| Reshape | (1, 64, 64, 2) | 0 |
| DensityMatrix | (1, 64, 64) | 0 |
| Expectation | (1, 4096) | 0 |

### SVAE MODEL

Tabel 7.10: SVAE parameters summary.

| **Parameter** | **Value** | **Description** |
|---|---|---|
| Shots | 100,000 | Number of measurement shots |
| Beta | 0.819 | Hyperparameter for regularization |
| Number of Steps | 100 | Training steps per epoch |
| Number of Epochs | 5 | Total epochs for training |
| Learning Rate | $1 \times 10^{-3}$ | Initial learning rate |
| Number of Workers | 4 | For data loading |
| Shuffle | False | Data shuffling disabled |
| Input Size | $4 \cdot n$ | Proportional to the number of qubits ($n$) |
| Hidden Size | $20 \cdot n$ | Proportional to the number of qubits ($n$) |
| Output (Latent) Size | $2 \cdot 2^n$ | Proportional to the number of qubits ($n$) |
| Alpha | 1 | Scaling factor for loss terms |
| Model Recovery | False | No recovery of previous models |

The SVAE model relies on a set of key parameters that govern its training and validation processes, as summarized in Tables 7.10 and 7.11. Table 7.10 lists the common parameters applied across all SVAE model tests, such as the total number of shots (100,000), hyperparameters like beta (0.819), learning rate ($1 \times 10^{-3}$), and architectural choices that scale proportionally with the number of qubits, including input size ($4 \cdot n$), hidden size ($20 \cdot n$), and output (latent) size ($2 \cdot 2^n$). Other important parameters include the number of training steps per epoch (100), total epochs (5), number of data-loading workers (4), and disabling of data shuffling during training.

Table 7.11 further details how training and validation are adapted for different numbers of qubits. The batch size for training gradually increases from 100 for 3

Tabel 7.11: Parameters for training and validation across different numbers of qubits.

| Qubits | Batch Size for Training | Validation Samples |
|---|---|---|
| 3 | 100 | 20,000 |
| 4 | 300 | 100,000 |
| 5 | 600 | $4^5 \cdot 500 = 512,000$ |
| 6 | 600 | $4^6 \cdot 500 = 2,048,000$ |
| 7 | 600 | $4^7 \cdot 500 = 8,192,000$ |
| 8 | 1,000 | $4^8 \cdot 500 = 32,768,000$ |

qubits up to 1,000 for 8 qubits, while the validation sample sizes grow exponentially with qubit number, following the pattern $4^n \times 500$, reaching over 32 million samples for 8 qubits. This scaling reflects the combinatorial complexity of quantum state representations and ensures the model is validated with sufficient data to capture the increasing state space.

## 7.6. CONCLUSION OF THIS CHAPTER

In this chapter, we demonstrated that memristor-compatible neural networks provide an effective, energy-efficient route to scalable quantum state tomography (QST). We first formalized the QST problem and contrasted two practical data-acquisition schemes (expectation-value based, M1, versus eigenstate-probability based, M2), highlighting how the choice of measurement bases and shot budget directly impacts reconstruction fidelity and runtime. We then evaluated a suite of architectures: FCN, CNN, RNN, RBM, CGAN, Transformer, and SVAE, under identical datasets and metrics.

Our results show that CNNs deliver the best accuracy–latency trade-off for moderate system sizes, particularly under M1, while the spiking variational autoencoder (SVAE) achieves the most favorable scaling with qubit number and measurement sparsity. The SVAE's probabilistic latent representation enables robust reconstructions for noisy, mixed states and sustains high fidelities from 3 to 8 qubits with substantially reduced computation time, establishing it as a promising backbone for larger systems.

Crucially, we proposed these algorithmic advances with in-memory computing (CiM) based on memristor crossbars, translating dense matrix–vector operations into analog current summations. Using the forming-free OxReRAM devices developed previously, we showed that weight programming granularity and multi-bit retention are sufficient to realize the required linear operations and on-chip inference with orders-of-magnitude lower energy than digital baselines. Together with the cryo-memristor results of Chapter 6, this establishes a clear hardware path toward co-locating QST inference near quantum processors, reducing I/O and thermal overheads.

**Limitations and Outlook.** Two practical gaps remain: (i) calibration of device non-idealities (IR drops, conductance drift, and read noise) to preserve network accuracy at scale, and (ii) co-design of measurement policies with the model (active or adaptive tomography) to minimize shots under fixed fidelity targets. Addressing these with closed-loop training, noise-aware quantization of weights/activations, and on-chip error compensation will be key for extending beyond 8 qubits. Looking forward, integrating SVAE-based inference with cryogenic CiM arrays and task-tailored measurement bases offers a concrete route to real-time, resource-frugal QST for larger quantum systems.

# 8

# CONCLUSIONS AND OUTLOOKS

## 8.1. RESTATING THE PROBLEM

Modern information processing is increasingly constrained by the energy, latency, and physical limits of conventional von Neumann architectures. This thesis explored *memristive devices and systems* across two tightly coupled axes: (i) materials and device engineering of (cryo-)ReRAM for reliable multi-level and ultra-stable operation, and (ii) algorithm–hardware co-design for energy-efficient neural inference, with quantum state tomography (QST) as a demanding, real-world benchmark where data movement, noise, and model scaling are critical.

## 8.2. RESEARCH QUESTIONS AND ANSWERS

*Can we realize forming-free, analog/multi-bit $HfO_2$-based ReRAM with adequate stability for compute-in-memory (CiM)?*
**Answer.** Yes. We introduced PdNeuRAM devices based on Pd/$HfO_2$ stacks that exhibit forming-free switching, multi-bit conductance tuning, and cycle-to-cycle stability compatible with vector–matrix operations. The material stack and interface engineering reduce stochastic forming and improve linearity, enabling practical weight programming for analog inference.

*How do cryogenic conditions impact device physics and can cryo-memristors meet quantum-adjacent requirements (e.g., low noise, fine biasing)?*
**Answer.** We demonstrated highly reliable cryogenic memristors with low conductance noise and fine-grained, stable levels suitable for sub-milliwatt operation at cryo temperatures. These characteristics support applications such as spin-qubit biasing and proximal analog compute, indicating a feasible path to near-quantum cryo-CMOS/CiM co-integration.

*Do memristor-compatible neural networks offer an energy-efficient, scalable route to QST?*
**Answer.** Yes. Among tested models (FCN, CNN, RBM, CGAN, Transformer, SVAE), a spiking VAE (SVAE) maintained robust fidelities under shot-limited and noisy regimes while mapping efficiently to CiM primitives. Convolutional baselines offered

favorable accuracy–latency trade-offs for moderate sizes, whereas SVAE scaled best as qubit count and measurement sparsity increased.

*What are the key co-design levers to close the model–hardware gap?*
**Answer.** Three levers proved essential: (i) noise-aware training that internalizes device non-idealities (IR-drop, drift, read noise), (ii) quantization strategies aligned with device conductance granularity, and (iii) adaptive measurement policies that co-optimize shots and model uncertainty (active tomography).

## 8.3. SYNTHESIS AND PERSPECTIVE

This thesis establishes a vertically integrated pathway: from interface-engineered, forming-free ReRAM and low-noise cryo-memristors, to neural architectures that are naturally decomposable into crossbar-friendly kernels, and finally to a QST workflow where *computation moves to the data*, near the quantum hardware and at cryogenic temperatures. In broader perspective, these results advance a unifying view: *memristive analog compute is most compelling where data locality, low-precision tolerance, and energy proportionality dominate system-level value.* QST is emblematic of such workloads, but the same principles extend to sensor fusion, edge inference, and cryo-control loops in quantum experiments.

## 8.4. MAIN CONTRIBUTIONS

1. Process/Integration contribution (liq-Si). We introduce a BEOL-compatible laser-induced liquid-phase–crystallized silicon (*liq-Si*) actuator that delivers highly localized, sub-$\mu$s thermal/electrical pulses for on-chip anneal, write-verify acceleration, and drift re-centering in Pd/$HfO_2$ ReRAM crossbars, improving uniformity and calibration without violating backend thermal budgets and remaining compatible with cryogenic integration.

2. Materials/device contribution: a forming-free, multi-bit Pd/$HfO_2$ ReRAM stack with improved linearity and stability for analog inference weights.

3. Cryogenic device contribution: demonstration of reliable, low-noise cryo-memristors tailored to quantum-adjacent operating points (fine biasing, stable multilevels).

4. Algorithmic contribution: benchmarking of NN families for QST, identifying CNNs for mid-scale efficiency and a spiking VAE (SVAE) for best scaling under noise/shot limits.

## 8.5. LIMITATIONS

While promising, the approach depends on robust calibration to mitigate IR-drop and conductance drift across large arrays. End-to-end system fidelity remains sensitive to peripheral non-idealities (DAC/ADC linearity, sense noise) and to drift over time/temperature. At scale, array tiling and interconnect overheads can

erode theoretical gains without careful architecture design. Finally, adaptive/active measurement in QST introduces control-loop complexity that must be co-verified with the physical experiment.

## 8.6. OUTLOOK AND FUTURE WORK

### (1) DEVICE AND MATERIALS

- Interface control: Systematic exploration of oxygen vacancy profiles (e.g., controlled scavenging layers, barrier insertions) to tune nonlinearity and retention.
- Drift/noise physics: Temperature-dependent studies and compact models linking microstructure to conductance drift and $1/f$ noise for design-time prediction.

### (2) CIRCUITS AND ARCHITECTURE

- Peripheral calibration: On-chip routines for write-verify, read-averaging, and background re-centering; lightweight ADC/DAC linearization.
- Array scaling: Tiled crossbar fabrics with hierarchical interconnect; co-simulation of latency/energy vs. fidelity under realistic non-idealities.

### (3) CRYOGENIC INTEGRATION

- Cryo-CMOS co-design: Joint optimization of biasing networks, sense amplifiers, and thermal budgets compatible with dilution refrigerator stages.
- Proximal compute: Validate closed-loop routines (state estimation, bias updates) with qubit hardware in the loop to quantify I/O and duty-cycle benefits.



### (4) ALGORITHMS FOR TOMOGRAPHY

- Active tomography: Integrate uncertainty-aware acquisition that adapts measurement bases and shot allocation online; analyze theoretical sample complexity.
- Noise-aware learning: Train with calibrated device noise/drift models; exploit Bayesian/spiking latent spaces (SVAE) to maintain fidelity at low precision.

### (5) BENCHMARKS AND TOOLING

- Open pipelines: Release datasets, quantization/noise injectors, and hardware-in-the-loop scripts to enable reproducible cross-lab comparisons.
- Cross-domain tasks: Extend beyond QST (e.g., cryo-control, sensor fusion), testing generality of the co-design principles.

## 8.7. CLOSING REMARKS

By aligning device physics, circuit calibration, and model design, this work demonstrates that memristive CiM can deliver tangible system-level benefits on a task that genuinely stresses fidelity, noise-robustness, and scalability. The results suggest a credible route to *proximal, energy-frugal intelligence*, from quantum labs to edge platforms, where moving compute towards the data is not only efficient but increasingly necessary.

# BIBLIOGRAPHY


[1] M. M. Waldrop. “The chips are down for Moore’s law”. In: *Nature News* 530.7589 (2016), p. 144.

[2] P. E. Ceruzzi. *A history of modern computing*. MIT press, 2003.

[3] J. L. Hennessy en D. A. Patterson. “A new golden age for computer architecture”. In: *Communications of the ACM* 62.2 (2019), p. 48–60.

[4] K. Asanovic, R. Bodik, B. Catanzaro, J. Gebis, P. Husbands, K. Keutzer, D. Patterson, W. Plishker, J. Shalf en S. W. Williams. “The landscape of parallel computing research: A view from berkeley”. In: (2006).

[5] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *e.a.* “Quantum supremacy using a programmable superconducting processor”. In: *Nature* 574.7779 (2019), p. 505–510.

[6] J. Backus. “Can programming be liberated from the von Neumann style? A functional style and its algebra of programs”. In: *Communications of the ACM* 21.8 (1978), p. 613–641.

[7] L. Atzori, A. Iera en G. Morabito. “The internet of things: A survey”. In: *Computer networks* 54.15 (2010), p. 2787–2805.

[8] M. Chen, S. Mao en Y. Liu. “Big data: A survey”. In: *Mobile networks and applications* 19 (2014), p. 171–209.

[9] C. Mead en M. Ismail. *Analog VLSI implementation of neural systems*. Deel 80. Springer Science & Business Media, 2012.

[10] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *e.a.* “A million spiking-neuron integrated circuit with a scalable communication network and interface”. In: *Science* 345.6197 (2014), p. 668–673.

[11] G. Indiveri en S.-C. Liu. “Memory and information processing in neuromorphic systems”. In: *Proceedings of the IEEE* 103.8 (2015), p. 1379–1397.

[12] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska en G. Indiveri. “A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses”. In: *Frontiers in neuroscience* 9 (2015), p. 141.

[13] D. Kuzum, R. G. Jeyasingh, B. Lee en H.-S. P. Wong. “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing”. In: *Nano letters* 12.5 (2012), p. 2179–2186.

[14] D. B. Strukov, G. S. Snider, D. R. Stewart en R. S. Williams. “The missing memristor found”. In: *nature* 453.7191 (2008), p. 80–83.

[15] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev en D. B. Strukov. “Training and operation of an integrated neuromorphic network based on metal-oxide memristors”. In: *Nature* 521.7550 (2015), p. 61–64.

[16] H.-S. P. Wong en S. Salahuddin. “Memory leads the way to better computing”. In: *Nature nanotechnology* 10.3 (2015), p. 191–194.

[17] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum en H.-S. P. Wong. “An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation”. In: *IEEE Transactions on Electron Devices* 58.8 (2011), p. 2729–2737.

[18] J. Haah, A. W. Harrow, Z. Ji, X. Wu en N. Yu. “Sample-optimal tomography of quantum states”. In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, p. 913–925.

[19] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker en J. Eisert. “Quantum state tomography via compressed sensing”. In: *Physical review letters* 105.15 (2010), p. 150401.

[20] H.-Y. Huang, R. Kueng en J. Preskill. “Predicting many properties of a quantum system from very few measurements”. In: *Nature Physics* 16.10 (2020), p. 1050–1057.

[21] P.-A. Mouny, R. Dawant, P. Dufour, M. Valdenaire, S. Ecoffey, M. Pioro-Ladrière, Y. Beilliard en D. Drouin. “Towards scalable cryogenic quantum dot biasing using memristor-based DC sources”. In: *Cryogenics* 142 (2024), p. 103910.

[22] J. Preskill. “Quantum computing in the NISQ era and beyond”. In: *Quantum* 2 (2018), p. 79.

[23] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha *e.a.* “Equivalent-accuracy accelerated neural-network training using analogue memory”. In: *Nature* 558.7708 (2018), p. 60–67.

[24] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko en G. Carleo. “Neural-network quantum state tomography”. In: *Nature physics* 14.5 (2018), p. 447–450.

[25] S. Yin, Z. Jiang, M. Kim, T. Gupta, M. Seok en J.-S. Seo. “Vesti: Energy-efficient in-memory computing accelerator for deep neural networks”. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 28.1 (2019), p. 48–61.

[26] J. A. Rogers, T. Someya en Y. Huang. “Materials and mechanics for stretchable electronics”. In: *science* 327.5973 (2010), p. 1603–1607.

[27] D.-H. Kim, N. Lu, R. Ma, Y.-S. Kim, R.-H. Kim, S. Wang, J. Wu, S. M. Won, H. Tao, A. Islam *e.a.* “Epidermal electronics”. In: *science* 333.6044 (2011), p. 838–843.

[28] Y. Chen. “Nanofabrication by electron beam lithography and its applications: A review”. In: *Microelectronic Engineering* 135 (2015), p. 57–72.

[29] S. M. Sze, Y. Li en K. K. Ng. *Physics of semiconductor devices*. John wiley & sons, 2021.

[30] A. Krasheninnikov en K. Nordlund. “Ion and electron irradiation-induced effects in nanostructured materials”. In: *Journal of applied physics* 107.7 (2010).

[31] M. Huth. “Focused electron beam induced deposition-principles and applications”. In: *Proceedings of Beilstein Symposium Functional Nanosciences, Bozen, Italy*. 2010, p. 193–211.

[32] D. Kudithipudi, C. Schuman, C. M. Vineyard, T. Pandit, C. Merkel, R. Kubendran, J. B. Aimone, G. Orchard, C. Mayr, R. Benosman *e.a.* “Neuromorphic computing at scale”. In: *Nature* 637.8047 (2025), p. 801–812.

[33] J. Park, A. Kumar, Y. Zhou, S. Oh, J.-H. Kim, Y. Shi, S. Jain, G. Hota, E. Qiu, A. L. Nagle *e.a.* “Multi-level, forming and filament free, bulk switching trilayer RRAM for neuromorphic computing at the edge”. In: *Nature Communications* 15.1 (2024), p. 3492.

[34] I.-W. Chen en J. Lee. *Ultrathin film resistive memory devices*. US Patent App. 14/507,957. 2016.

[35] Y.-B. Wang, T.-C. Chang, S.-K. Lin, P.-Y. Wu, Y.-C. Zhang, Y.-F. Tan, W.-C. Chen, C.-W. Wu, S.-Y. Chou, K.-J. Zhou *e.a.* “Forming-Free HfO 2-Based Resistive Random Access Memory by X-ray Irradiation”. In: *IEEE Transactions on Electron Devices* 69.12 (2022), p. 6705–6709.

[36] Z. Wang, R. Yang, H.-M. Huang, H.-K. He, J. Shaibo en X. Guo. “Electroforming-Free Artificial Synapses Based on Proton Conduction in $\alpha$-MoO3 Films”. In: *Advanced Electronic Materials* 6.4 (2020), p. 1901290.

[37] X. Ding, X. Yu, Z. Lan, J. Li, S. Zhou, C. Lee en Y. Zhao. “Forming-Free HfO x-Based Resistive Memory With Improved Uniformity Achieved by the Thermal Annealing-Induced Self-Doping of Ge”. In: *IEEE Transactions on Electron Devices* 70.4 (2023), p. 1671–1675.

[38] W. Kim, A. Hardtdegen, C. Rodenbücher, S. Menzel, D. J. Wouters, S. Hoffmann-Eifert, D. Buca, R. Waser en V. Rana. “Forming-free metal-oxide ReRAM by oxygen ion implantation process”. In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2016, p. 4–4.

[39] W. Kim, D. J. Wouters, S. Menzel, C. Rodenbücher, R. Waser en V. Rana. “Lowering forming voltage and forming-free behavior of ta 2 o 5 reram devices”. In: *2016 46th European Solid-State Device Research Conference (ESSDERC)*. IEEE. 2016, p. 164–167.

[40] E. Wu, T. Ando en P. Jamison. “Reversing a decades-long scaling law of dielectric breakdown for ReRAM forming voltage reduction-Modeling competition among defect generation and annihilation processes”. In: *2023 International Electron Devices Meeting (IEDM)*. IEEE. 2023, p. 1–4.

[41] B. Boro, R. Parmar en G. Trivedi. “Vector-Matrix Multiplier Architecture for In-Memory Computing Applications with RRAM Arrays”. In: *IEEE Transactions on Nanotechnology* (2025).

[42] A. Zwerver, T. Krähenmann, T. Watson, L. Lampert, H. C. George, R. Pillarisetty, S. Bojarski, P. Amin, S. Amitonov, J. Boter *e.a.* “Qubits made by advanced semiconductor manufacturing”. In: *Nature Electronics* 5.3 (2022), p. 184–190.

[43] L. M. Vandersypen, H. Bluhm, J. S. Clarke, A. S. Dzurak, R. Ishihara, A. Morello, D. J. Reilly, L. R. Schreiber en M. Veldhorst. “Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent”. In: *npj Quantum Information* 3.1 (2017), p. 34.

[44] P.-A. Mouny, Y. Beilliard, S. Graveline, M.-A. Roux, A. El Mesoudy, R. Dawant, P. Gliech, S. Ecoffey, F. Alibart, M. Pioro-Ladrière *e.a.* “Memristor-based cryogenic programmable DC sources for scalable in situ quantum-dot control”. In: *IEEE Transactions on Electron Devices* 70.4 (2023), p. 1989–1995.

[45] L. Petit, H. Eenink, M. Russ, W. Lawrie, N. Hendrickx, S. Philips, J. Clarke, L. Vandersypen en M. Veldhorst. “Universal quantum logic in hot silicon qubits”. In: *Nature* 580.7803 (2020), p. 355–359.

[46] S. Alam, M. S. Hossain, S. R. Srinivasa en A. Aziz. “Cryogenic memory technologies”. In: *Nature Electronics* 6.3 (2023), p. 185–198.

[47] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh en E. Eleftheriou. “Memory devices and applications for in-memory computing”. In: *Nature nanotechnology* 15.7 (2020), p. 529–544.

[48] J. Carrasquilla, G. Torlai, R. G. Melko en L. Aolita. “Reconstructing quantum states with generative models”. In: (okt 2018). DOI: 10.1038/s42256-019-0028-1. URL: http://arxiv.org/abs/1810.10584%20http://dx.doi.org/10.1038/s42256-019-0028-1.

[49] H. Ma, D. Dong, I. R. Petersen, C.-J. Huang en G.-Y. Xiang. “A comparative study on how neural networks enhance quantum state tomography”. In: (2021), p. 1–12. arXiv: 2111.09504. URL: http://arxiv.org/abs/2111.09504.

[50] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang en H. Qian. “Fully hardware-implemented memristor convolutional neural network”. In: *Nature* 577.7792 (2020), p. 641–646.

[51] T. Schmale, M. Reh en M. Gärttner. “Efficient quantum state tomography with convolutional neural networks”. In: *npj Quantum Information* 8.1 (2022), p. 115.

[52] Y. Quek, S. Fort en H. K. Ng. “Adaptive quantum state tomography with neural networks”. In: *npj Quantum Information* 7.1 (2021), p. 105.

[53] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *e.a.* “Loihi: A neuromorphic manycore processor with on-chip learning”. In: *Ieee Micro* 38.1 (2018), p. 82–99.

[54] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. v. Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud *e.a.* "Neuromorphic silicon neuron circuits". In: *Frontiers in neuroscience* 5 (2011), p. 73.

[55] G. Carleo en M. Troyer. "Solving the quantum many-body problem with artificial neural networks". In: *Science* 355.6325 (2017), p. 602–606.

[56] C. Mead. "Neuromorphic electronic systems". In: *Proceedings of the IEEE* 78.10 (1990), p. 1629–1636.

[57] S. Choi, T. Moon, G. Wang en J. J. Yang. "Filament-free memristors for computing". In: *Nano Convergence* 10.1 (2023), p. 58.

[58] L. Chua. "Memristor-the missing circuit element". In: *IEEE Transactions on circuit theory* 18.5 (1971), p. 507–519.

[59] Y. Xiao, B. Jiang, Z. Zhang, S. Ke, Y. Jin, X. Wen en C. Ye. "A review of memristor: material and structure design, device performance, applications and prospects". In: *Science and Technology of Advanced Materials* 24.1 (2023), p. 2162323.

[60] S. Chen, Z. Yang, H. Hartmann, A. Besmehn, Y. Yang en I. Valov. "Electrochemical ohmic memristors for continual learning". In: *Nature Communications* 16.1 (2025), p. 2348.

[61] M. Wuttig en N. Yamada. "Phase-change materials for rewriteable data storage". In: *Nature materials* 6.11 (2007), p. 824–832.

[62] I. Boybat, M. Le Gallo, S. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian en E. Eleftheriou. "Neuromorphic computing with multi-memristive synapses". In: *Nature communications* 9.1 (2018), p. 2514.

[63] D. Loke, T. Lee, W. Wang, L. Shi, R. Zhao, Y. Yeo, T. Chong en S. Elliott. "Breaking the speed limits of phase-change memory". In: *Science* 336.6088 (2012), p. 1566–1569.

[64] M. Le Gallo en A. Sebastian. "An overview of phase-change memory device physics". In: *Journal of Physics D: Applied Physics* 53.21 (2020), p. 213002.

[65] D. Ielmini, S. Lavizzari, D. Sharma en A. L. Lacaita. "Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation". In: *2007 IEEE International Electron Devices Meeting*. IEEE. 2007, p. 939–942.

[66] S. Kohara, K. Kato, S. Kimura, H. Tanaka, T. Usuki, K. Suzuya, H. Tanaka, Y. Moritomo, T. Matsunaga, N. Yamada *e.a.* "Structural basis for the fast phase change of Ge2Sb2Te5: Ring statistics analogy between the crystal and amorphous states". In: *Applied physics letters* 89.20 (2006).

[67] M. Trentzsch, S. Flachowsky, R. a. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck *e.a.* "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs". In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2016, p. 11–5.

[68] H. Mulaosmanovic, E. T. Breyer, S. Dünkel, S. Beyer, T. Mikolajick en S. Slesazeck. “Ferroelectric field-effect transistors based on HfO2: a review”. In: *Nanotechnology* 32.50 (2021), p. 502002.

[69] T. Böscke, J. Müller, D. Braeuhaus, U. Schröder en U. Böttger. “Ferroelectricity in hafnium oxide thin films”. In: *Applied Physics Letters* 99.10 (2011).

[70] J. Y. Kim, M.-J. Choi en H. W. Jang. “Ferroelectric field effect transistors: Progress and perspective”. In: *APL Materials* 9.2 (2021).

[71] K. Roy, A. Jaiswal en P. Panda. “Towards spike-based machine intelligence with neuromorphic computing”. In: *Nature* 575.7784 (nov 2019), p. 607–617. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1677-2. URL: https://doi.org/10.1038/s41586-019-1677-2.

[72] G. Srinivasan, A. Sengupta en K. Roy. “Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning”. In: *Scientific reports* 6.1 (2016), p. 29545.

[73] L. Chen, T.-Y. Wang, Y.-W. Dai, M.-Y. Cha, H. Zhu, Q.-Q. Sun, S.-J. Ding, P. Zhou, L. Chua en D. W. Zhang. “Ultra-low power Hf 0.5 Zr 0.5 O 2 based ferroelectric tunnel junction synapses for hardware neural network applications”. In: *Nanoscale* 10.33 (2018), p. 15826–15833.

[74] E. Hua, C. Glessi, C. W. Hagen en R. Ishihara. “"Novel Approach to a-Si Formation Using Eletron Beam Induced Processing with Liquid Heavy Polysilanes"”. Manuscript in submission to APL Materials. 2025.

[75] T. Shimoda, Y. Matsuki, M. Furusawa, T. Aoki, I. Yudasaka, H. Tanaka, H. Iwasawa, D. Wang, M. Miyasaka en Y. Takeuchi. “Solution-processed silicon films and transistors”. In: *Nature* 440.7085 (2006), p. 783–786.

[76] J. Zhang, M. Trifunovic, M. Van Der Zwan, H. Takagishi, R. Kawajiri, T. Shimoda, C. Beenakker en R. Ishihara. “Single-grain Si thin-film transistors on flexible polyimide substrate fabricated from doctor-blade coated liquid-Si”. In: *Applied Physics Letters* 102.24 (2013).

[77] M. Trifunovic, T. Shimoda en R. Ishihara. “Solution-processed polycrystalline silicon on paper”. In: *Applied Physics Letters* 106.16 (2015).

[78] M. Trifunovic, P. M. Sberna, T. Shimoda en R. Ishihara. “Solution-based polycrystalline silicon transistors produced on a paper substrate”. In: *npj Flexible Electronics* 1.1 (2017), p. 12.

[79] N. Farid, A. Brunton, P. Rumsby, S. Monaghan, R. Duffy, P. Hurley, M. Wang, K.-L. Choy en G. M. O’Connor. “Femtosecond laser-induced crystallization of amorphous silicon thin films under a thin molybdenum layer”. In: *ACS Applied Materials & Interfaces* 13.31 (2021), p. 37797–37808.

[80] W. F. Van Dorp, B. Van Someren, C. W. Hagen, P. Kruit en P. A. Crozier. “Approaching the resolution limit of nanometer-scale electron beam-induced deposition”. In: *Nano letters* 5.7 (2005), p. 1303–1307.

[81] S. Randolph, J. Fowlkes en P. Rack. “Focused, nanoscale electron-beam-induced deposition and etching”. In: *Critical reviews in solid state and materials sciences* 31.3 (2006), p. 55–89.

[82] D. Hess, D. Graves, M. Hitchman en K. Jensen. “Chemical Vapor Deposition: Principles and Applications”. In: *Chemical Vapour Deposition: Principles and Applications* (1993).

[83] M. Ohring. *Materials science of thin films: depositon and structure*. Academic press, 2002.

[84] S. M. George. “Atomic layer deposition: an overview”. In: *Chemical reviews* 110.1 (2010), p. 111–131.

[85] R. L. Puurunen. “A short history of atomic layer deposition: Tuomo Suntola’s atomic layer epitaxy”. In: *Chemical Vapor Deposition* 20.10-11-12 (2014), p. 332–344.

[86] M. R. Baklanov, C. Adelmann, L. Zhao en S. De Gendt. “Advanced interconnects: materials, processing, and reliability”. In: *ECS Journal of Solid State Science and Technology* 4.1 (2015), Y1–Y4.

[87] I. Utke, P. Hoffmann en J. Melngailis. “Gas-assisted focused electron beam and ion beam processing and fabrication”. In: *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 26.4 (2008), p. 1197–1276.

[88] M. Huth, F. Porrati en O. V. Dobrovolskiy. “Focused electron beam induced deposition meets materials science”. In: *Microelectronic Engineering* 185 (2018), p. 9–28.

[89] S. Uchikoga en N. Ibaraki. “Low temperature poly-Si TFT-LCD by excimer laser anneal”. In: *Thin Solid Films* 383.1-2 (2001), p. 19–24.

[90] C. Becker, D. Amkreutz, T. Sontheimer, V. Preidel, D. Lockau, J. Haschke, L. Jogschies, C. Klimm, J. J. Merkel, P. Plocica *e.a.* “Polycrystalline silicon thin-film solar cells: Status and perspectives”. In: *Solar Energy Materials and Solar Cells* 119 (2013), p. 112–123.

[91] J. I. Goldstein, D. E. Newbury, J. R. Michael, N. W. Ritchie, J. H. J. Scott en D. C. Joy. *Scanning electron microscopy and X-ray microanalysis*. springer, 2017.

[92] H. Bartzsch, D. Glöß, B. Böcher, P. Frach en K. Goedicke. “Properties of SiO2 and Al2O3 films for electrical insulation applications deposited by reactive pulse magnetron sputtering”. In: *Surface and Coatings Technology* 174 (2003), p. 774–778.

[93] S. M. Sze, Y. Li en K. K. Ng. *Physics of semiconductor devices*. John wiley & sons, 2021.

[94] E. Hua, H. Abunahla, G. Gaydadjiev, S. Hamdioui en R. Ishihara. “Multi-level forming-free HfO 2-based ReRAM for energy-efficient computing”. In: *2024 Device Research Conference (DRC)*. IEEE. 2024, p. 1–2.

[95] E. Hua, T. Spyrou, M. Ahmadi, A. M. Syed, H. Xun, L. Braic, E. van der Veer, N. Elatab, A. Gebregiorgis, G. Gaydadjiev *e.a.* “PdNeuRAM: Forming-Free, Multi-Bit Pd/HfO2 ReRAM for Energy-Efficient Computing”. In: *arXiv preprint arXiv:2505.22789* (2025).

[96] E. Hua, H. Abunahla en R. Ishihara. “Resistive Random Access Memory for performing forming-free resistive switching and method of manufacturing the memory”. Patent NL2036647B1. T. Delft. (Granted). 2025. URL: https://patents.google.com/patent/NL2036647B1.

[97] E. Hua, T. Spyrou, M. Ahmadi, A. M. Syed, H. Xun, L. Braic, E. van der Veer, N. Elatab, A. Gebregiorgis, G. Gaydadjiev, B. Noheda, S. Hamdioui, R. Ishihara en H. Abunahla. “PdNeuRAM: Forming-Free, Multi-Bit Pd/$HfO_2$ ReRAM for Energy-Efficient Computing”. Nature Communications Engineering, in revision. 2025.

[98] S. Hamdioui, L. Xie, H. A. Du Nguyen, M. Taouil, K. Bertels, H. Corporaal, H. Jiao, F. Catthoor, D. Wouters, L. Eike *e.a.* “Memristor based computation-in-memory architecture for data-intensive applications”. In: *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2015, p. 1718–1725.

[99] C.-Y. Tsai, Y.-W. Ting en K.-C. Huang. *Structure and method for a forming free resistive random access memory with multi-level cell*. US Patent 9,053,781. 2015.

[100] Y.-S. Chen, H.-Y. Lee, P.-S. Chen, T.-Y. Wu, C.-C. Wang, P.-J. Tzeng, F. Chen, M.-J. Tsai en C. Lien. “An ultrathin forming-free HfOx resistance memory with excellent electrical performance”. In: *IEEE electron device letters* 31.12 (2010), p. 1473–1475.

[101] Y.-J. Huang, S. C. Pan en S.-C. Lee. *RRAM devices*. US Patent 9,525,008. 2016.

[102] T. Stecconi, R. Guido, L. Berchialla, A. La Porta, J. Weiss, Y. Popoff, M. Halter, M. Sousa, F. Horst, D. Dávila *e.a.* “Filamentary TaOx/HfO2 ReRAM Devices for Neural Networks Training with Analog In-Memory Computing”. In: *Advanced electronic materials* 8.10 (2022), p. 2200448.

[103] A. Wedig, M. Luebben, D.-Y. Cho, M. Moors, K. Skaja, V. Rana, T. Hasegawa, K. K. Adepalli, B. Yildiz, R. Waser *e.a.* “Nanoscale cation motion in TaO x, HfO x and TiO x memristive systems”. In: *Nature nanotechnology* 11.1 (2016), p. 67–74.

[104] P. Nukala, M. Ahmadi, Y. Wei, S. De Graaf, E. Stylianidis, T. Chakrabortty, S. Matzen, H. W. Zandbergen, A. Björling, D. Mannix *e.a.* “Reversible oxygen migration and phase transitions in hafnia-based ferroelectric devices”. In: *Science* 372.6542 (2021), p. 630–635.

[105] D. F. Falcone, V. Clerico, W. Choi, T. Stecconi, F. Horst, L. Begon-Lours, M. Galetta, A. La Porta, N. Garg, F. Alibart *e.a.* “All-in-One Analog AI Accelerator: On-Chip Training and Inference with Conductive-Metal-Oxide/HfOx ReRAM Devices”. In: *arXiv preprint arXiv:2502.04524* (2025).

[106] K. Beckmann, J. Holt, H. Manem, J. Van Nostrand en N. C. Cady. “Nanoscale hafnium oxide rram devices exhibit pulse dependent behavior and multi-level resistance capability”. In: *Mrs Advances* 1.49 (2016), p. 3355–3360.

[107] K. Nakayama, M. Takata, T. Kasai, A. Kitagawa en J. Akita. “Pulse number control of electrical resistance for multi-level storage based on phase change”. In: *Journal of Physics D: Applied Physics* 40.17 (2007), p. 5061.

[108] C. Xu, D. Niu, N. Muralimanohar, N. P. Jouppi en Y. Xie. “Understanding the trade-offs in multi-level cell ReRAM memory design”. In: *Proceedings of the 50th Annual Design Automation Conference*. 2013, p. 1–6.

[109] F. A. Ikraiam. “Investigating the Effects of Ionizing Radiation on CV Characteristics of SOI-MOS Capacitors”. In: *Proceedings of the 14th Arabic Conference of the Arab Union for Astronomy and Space Sciences: AUASS-CONF23, 13-16 November 2023, Sharjah, United Arab Emirates*. Deel 420. Springer Nature. 2025, p. 266.

[110] B. Traoré, P. Blaise, E. Vianello, L. Perniola, B. De Salvo en Y. Nishi. “HfO 2-based RRAM: Electrode effects, Ti/HfO 2 interface, charge injection, and oxygen (O) defects diffusion through experiment and ab initio calculations”. In: *IEEE Transactions on Electron Devices* 63.1 (2015), p. 360–368.

[111] A. Hardtdegen, H. Zhang en S. Hoffmann-Eifert. “Tuning the performance of Pt/HfO2/Ti/Pt ReRAM devices obtained from plasma-enhanced atomic layer deposition for HfO2 thin films”. In: *ECS Transactions* 75.6 (2016), p. 177.

[112] Q. Luo, X. Zhang, Y. Hu, T. Gong, X. Xu, P. Yuan, H. Ma, D. Dong, H. Lv, S. Long *e.a.* “Self-rectifying and forming-free resistive-switching device for embedded memory application”. In: *IEEE Electron Device Letters* 39.5 (2018), p. 664–667.

[113] Y. Sun, Z. Ma, D. Tan, Z. Shen, J. You, W. Li, L. Xu, K. Chen en D. Feng. “Hybrid channel induced forming-free performance in nanocrystalline-Si: H/a-SiNx: H resistive switching memory”. In: *Nanotechnology* 30.36 (2019), p. 365701.

[114] B. Chakrabarti, R. V. Galatage en E. M. Vogel. “Multilevel Switching in Forming-Free Resistive Memory Devices With Atomic Layer Deposited $HfTiO_x$ Nanolaminate”. In: *IEEE electron device letters* 34.7 (2013), p. 867–869.

[115] S. Kumar en C. Rath. “Oxygen Vacancy Mediated Stabilization of Cubic Phase at Room Temperature and Resistive Switching Effect in Sm-and Dy-Doped HfO2 Thin Film”. In: *physica status solidi (a)* 217.1 (2020), p. 1900756.

[116] P. Calka, M. Sowinska, T. Bertaud, D. Walczyk, J. Dabrowski, P. Zaumseil, C. Walczyk, A. Gloskovskii, X. Cartoixà, J. Suñé *e.a.* “Engineering of the Chemical Reactivity of the Ti/HfO2 Interface for RRAM: Experiment and Theory.” In: *ACS applied materials & interfaces* 6.7 (2014), p. 5056–5060.

[117] S. R. Bradley, K. P. McKenna en A. L. Shluger. “The behaviour of oxygen at metal electrodes in HfO2 based resistive switching devices”. In: *Microelectronic engineering* 109 (2013), p. 346–350.

[118] G. Niu, H.-D. Kim, R. Roelofs, E. Perez, M. A. Schubert, P. Zaumseil, I. Costina en C. Wenger. “Material insights of HfO2-based integrated 1-transistor-1-resistor resistive random access memory devices processed by batch atomic layer deposition”. In: *Scientific reports* 6.1 (2016), p. 28155.

[119] R. Viennet, H. Roussel, L. Rapenne, J. Deschanvres, H. Renevier, V. Jousseaume, E. Jalaguier en M. Proietti. “XAFS atomistic insight of the oxygen gettering in Ti/HfO 2 based OxRRAM”. In: *Physical Review Materials* 2.5 (2018), p. 055002.

[120] G. Orchard, A. Jayawant, G. K. Cohen en N. Thakor. “Converting static image datasets to spiking neuromorphic datasets using saccades”. In: *Frontiers in Neuroscience* 9 (2015). Article 437.

[121] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. D. Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner en D. Modha. “A Low Power, Fully Event-Based Gesture Recognition System”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[122] S. B. Shrestha en G. Orchard. “SLAYER: Spike layer error reassignment in time”. In: *Advances in Neural Information Processing Systems*. 2018, p. 1412–1421.

[123] N. Capron, P. Broqvist en A. Pasquarello. “Migration of oxygen vacancy in HfO2 and across the HfO2/ SiO2 interface: A first-principles investigation”. In: *Applied Physics Letters* 91.19 (2007).

[124] Z. Hou, X. Gong en Q. Li. “Al-induced reduction of the oxygen diffusion in HfO2: an ab initio study”. In: *Journal of Physics: Condensed Matter* 20.13 (2008), p. 135206.

[125] R. Öttking, S. Kupke, E. Nadimi, R. Leitsmann, F. Lazarevic, P. Plänitz, G. Roll, S. Slesazeck, M. Trentzsch en T. Mikolajick. “Defect generation and activation processes in HfO2 thin films: Contributions to stress-induced leakage currents”. In: *physica status solidi (a)* 212.3 (2015), p. 547–553.

[126] A. S. Foster, F. L. Gejo, A. Shluger en R. M. Nieminen. “Vacancy and interstitial defects in hafnia”. In: *Physical Review B* 65.17 (2002), p. 174117.

[127] R. Waser en M. Aono. “Nanoionics-based resistive switching memories”. In: *Nature materials* 6.11 (2007), p. 833–840.

[128] H. A. Tahini, A. Chroneos, S. C. Middleburgh, U. Schwingenschlögl en R. W. Grimes. “Ultrafast palladium diffusion in germanium”. In: *Journal of Materials Chemistry A* 3.7 (2015), p. 3832–3838.

[129] H. B. Do, Q. H. Luc, M. T. H. Ha, S. H. Huynh, T. A. Nguyen, Y. C. Lin en E. Y. Chang. “Study of the interface stability of the metal (Mo, Ni, Pd)/HfO2/AlN/InGaAs MOS devices”. In: *AIP Advances* 7.8 (2017).

[130] D. Yamashita en A. Ishizaki. “In situ measurements of change in work function of Pt, Pd and Au surfaces during desorption of oxygen by using photoemission yield spectrometer in air”. In: *Applied Surface Science* 363 (2016), p. 240–244.

[131] J. Giner en E. Lange. “Elektronenaustrittsspannungen von reinem und sauerstoffbedecktem Au, Pt und Pd auf Grund ihrer Voltaspannungen gegen Ag”. In: *Naturwissenschaften* 40.19 (1953), p. 506–506.

[132] R. Jaeckel en B. Wagner. “Photo-electric measurement of the work function of metals and its alteration after gas adsorption”. In: *Vacuum* 13.12 (1963), p. 509–511.

[133] X. Shen, W. Liu, X. Gao, Z. Lu, X. Wu en X. Gao. “Mechanisms of oxidase and superoxide dismutation-like activities of gold, silver, platinum, and palladium, and their alloys: a general way to the activation of molecular oxygen”. In: *Journal of the American Chemical Society* 137.50 (2015), p. 15882–15891.

[134] D. Gu, S. K. Dey en P. Majhi. “Effective work function of Pt, Pd, and Re on atomic layer deposited HfO2”. In: *Applied Physics Letters* 89.8 (2006).

[135] K.-Y. Tse en J. Robertson. “Control of Schottky barrier heights on high-k gate dielectrics for future complementary metal-oxide semiconductor devices”. In: *Physical review letters* 99.8 (2007), p. 086805.

[136] R. T. Tung. “Chemical bonding and Fermi level pinning at metal-semiconductor interfaces”. In: *Physical review letters* 84.26 (2000), p. 6078.

[137] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang en H. Qian. “Conduction mechanisms, dynamics and stability in ReRAMs”. In: *Microelectronic Engineering* 187 (2018), p. 121–133.

[138] C. Funck en S. Menzel. “Comprehensive model of electron conduction in oxide-based memristive devices”. In: *ACS Applied electronic materials* 3.9 (2021), p. 3674–3692.

[139] C. Bengel, A. Siemon, F. Cüppers, S. Hoffmann-Eifert, A. Hardtdegen, M. Von Witzleben, L. Hellmich, R. Waser en S. Menzel. “Variability-aware modeling of filamentary oxide-based bipolar resistive switching cells using SPICE level compact models”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.12 (2020), p. 4618–4630.

[140] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai en S. Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Red. door H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox en R. Garnett. Curran Associates, Inc., 2019, p. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[141] W. Gerstner. “Time structure of the activity in neural network models”. In: *Phys. Rev. E* 51 (1 1995), p. 738–758.

[142] E. Hua, S. Heemskerk, W. Braganc, H. Xun, A. Kaichouhi en R. Ishihara. “Ultra-Low-Noise Cryogenic Multilevel Memristors for Sub-100 $\mu$V Spin Qubit Biasing”. In: *Proceedings of the IEEE International Electron Devices Meeting (IEDM). Accepted*. San Francisco, USA: IEEE, 2025.

[143] J. J. Pla, K. Y. Tan, J. P. Dehollain, W. H. Lim, J. J. Morton, F. A. Zwanenburg, D. N. Jamieson, A. S. Dzurak en A. Morello. “High-fidelity readout and control of a nuclear spin qubit in silicon”. In: *Nature* 496.7445 (2013), p. 334–338.

[144] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol en D. M. Lucas. “High-fidelity quantum logic gates using trapped-ion hyperfine qubits”. In: *Physical review letters* 117.6 (2016), p. 060504.

[145] M. Malinowski, D. Allcock en C. Ballance. “How to wire a 1000-qubit trapped-ion quantum computer”. In: *PRX Quantum* 4.4 (2023), p. 040313.

[146] F. A. Zwanenburg, A. S. Dzurak, A. Morello, M. Y. Simmons, L. C. Hollenberg, G. Klimeck, S. Rogge, S. N. Coppersmith en M. A. Eriksson. “Silicon quantum electronics”. In: *Reviews of modern physics* 85.3 (2013), p. 961–1019.

[147] D. M. Zajac, A. J. Sigillito, M. Russ, F. Borjans, J. M. Taylor, G. Burkard en J. R. Petta. “Resonantly driven CNOT gate for electron spins”. In: *Science* 359.6374 (2018), p. 439–442.

[148] A. D. O’Connell, M. Ansmann, R. C. Bialczak, M. Hofheinz, N. Katz, E. Lucero, C. McKenney, M. Neeley, H. Wang, E. M. Weig *e.a.* “Microwave dielectric loss at single photon energies and millikelvin temperatures”. In: *Applied Physics Letters* 92.11 (2008).

[149] E. Hua, S. van Ommen, K. Y. Yu, J. van Leeuven, R. Bishnoi, H. Abunahla, S. Nur, S. Feld en R. Ishihara. “Neural Network Architectures for Scalable Quantum State Tomography: Benchmarking and Memristor-Based Acceleration”. In: *arXiv preprint arXiv:2507.23007* (2025).

[150] M. A. Nielsen en I. L. Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[151] P. W. Shor. “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”. In: *SIAM review* 41.2 (1999), p. 303–332.

[152] E. Farhi, J. Goldstone en S. Gutmann. “A quantum approximate optimization algorithm”. In: *arXiv preprint arXiv:1411.4028* (2014).

[153] S. Lloyd. “Universal quantum simulators”. In: *Science* 273.5278 (1996), p. 1073–1078.

[154] M. A. Nielsen en I. L. Chuang. “Quantum information and quantum computation”. In: *Cambridge: Cambridge University Press* 2.8 (2000), p. 23.

[155] M. Paris en J. Rehacek. *Quantum state estimation*. Deel 649. Springer Science & Business Media, 2004.

[156] D. F. James, P. G. Kwiat, W. J. Munro en A. G. White. “Measurement of qubits”. In: *Physical Review A* 64.5 (2001), p. 052312.

[157] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin en Y.-K. Liu. “Efficient quantum state tomography”. In: *Nature communications* 1.1 (2010), p. 149.

[158] D. Koutnỳ, L. Motka, Z. Hradil, J. Řeháček en L. L. Sánchez-Soto. “Neural-network quantum state tomography”. In: *Physical Review A* 106.1 (2022), p. 012409.

[159] A. A. Melnikov, H. Poulsen Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger en H. J. Briegel. “Active learning machine learns to create new quantum experiments”. In: *Proceedings of the National Academy of Sciences* 115.6 (2018), p. 1221–1226.

[160] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko en G. Carleo. “Neural-network quantum state tomography”. In: *Nature Physics* 14 (5 mei 2018), p. 447–450. ISSN: 17452481. DOI: 10.1038/s41567-018-0048-5.

[161] J. Carrasquilla, G. Torlai, R. G. Melko en L. Aolita. “Reconstructing quantum states with generative models”. In: *Nature Machine Intelligence* 1.3 (2019), p. 155–161.

[162] S. Lohani, B. T. Kirby, M. Brodsky, O. Danaci en R. T. Glasser. “Machine learning assisted quantum state estimation”. In: *Machine Learning: Science and Technology* 1.3 (2020), p. 035007.

[163] H. Ma, D. Dong, I. R. Petersen, C.-J. Huang en G.-Y. Xiang. “Neural networks for quantum state tomography with constrained measurements”. In: *Quantum Information Processing* 23.9 (2024), p. 317.

[164] S. Morawetz, I. J. De Vlugt, J. Carrasquilla en R. G. Melko. “U (1)-symmetric recurrent neural networks for quantum state reconstruction”. In: *Physical Review A* 104.1 (2021), p. 012401.

[165] A. Neville, C. Sparrow, R. Clifford, E. Johnston, P. M. Birchall, A. Montanaro en A. Laing. “Classical boson sampling algorithms with superior performance to near-term experiments”. In: *Nature Physics* 13.12 (2017), p. 1153–1157.

[166] S. Ahmed, C. Sánchez Muñoz, F. Nori en A. F. Kockum. “Quantum State Tomography with Conditional Generative Adversarial Networks”. In: *Physical Review Letters* 127.14 (2021), p. 1–8. ISSN: 10797114. DOI: 10.1103/PhysRevLett.127.140502. arXiv: 2008.03240.

[167] H. Ma, Z. Sun, D. Dong, C. Chen en H. Rabitz. “Attention-based transformer networks for quantum state tomography”. In: *arXiv preprint arXiv:2305.05433* (2023).

[168] H. Ma, Z. Sun, D. Dong en D. Gong. “Learning informative latent representation for quantum state tomography”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2025).

[169] A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo en S. Severini. “Learning hard quantum distributions with variational autoencoders”. In: *npj Quantum Information* 4.1 (2018), p. 28.

[170] C. Chen, Z. He, Z. Huang en H. Situ. “Reconstructing a quantum state with a variational autoencoder”. In: *International Journal of Quantum Information* 19.08 (2021), p. 2140005.

[171] M. A. Zidan, J. P. Strachan en W. D. Lu. “The future of electronics based on memristive systems”. In: *Nature electronics* 1.1 (2018), p. 22–29.

[172] D. Ielmini en H.-S. P. Wong. “In-memory computing with resistive switching devices”. In: *Nature electronics* 1.6 (2018), p. 333–343.

[173] M. Krenn, J. Landgraf, T. Foesel en F. Marquardt. *Artificial intelligence and machine learning for quantum technologies*. Jan 2023. DOI: 10.1103/PhysRevA.107.010101.

[174] S. in Yi, J. D. Kendall, R. S. Williams en S. Kumar. “Activity-difference training of deep neural networks using memristor crossbars”. In: *Nature Electronics* 6.1 (2023), p. 45–51. ISSN: 25201131. DOI: 10.1038/s41928-022-00869-w.

[175] D. P. Kingma en J. Ba. “Adam: A Method for Stochastic Optimization”. In: (dec 2014). URL: http://arxiv.org/abs/1412.6980.

[176] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn en W. D. Lu. “A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations”. In: *Nature Electronics* 2.July (2019), p. 290–299. ISSN: 2520-1131. URL: http://dx.doi.org/10.1038/s41928-019-0270-x.

[177] S. Dash. “Accurate & Energy-efficient ECG Classification using RRAM based DNN Architecture”. In: December 2020 (2021). URL: http://repository.tudelft.nl/..

[178] A. Patel, A. Gaikwad, T. Huang, A. F. Kockum en T. Abad. “Selective and efficient quantum state tomography for multi-qubit systems”. In: *arXiv preprint arXiv:2503.20979* (2025).

[179] S. T. Ahmad, A. Farooq en H. Shin. “Self-guided quantum state tomography for limited resources”. In: *Scientific Reports* 12.1 (2022), p. 5092.

# LIST OF PUBLICATIONS

Publications related to PhD research:


14. E. Hua en R. Ishihara. “Cryogenic Memristors for Scalable Qubit Control: Bridging Precision, Density, and Power Through Interface-Engineered $HfO_2$ Devices”. In: *Proceedings of the 2026 MRS Spring Meeting & Exhibit. Submitted*, Control ID: 4434113, Symposium EL01: Materials and Devices for Unconventional Computing. Honolulu, Hawai’i, USA: Materials Research Society, 2026

13. E. Hua, S. Heemskerk en R. Ishihara. “Sub-100 $\mu$V Cryogenic Biasing with Low-Noise Multilevel Memristors: Pt-Interface Tuning for Scalable Qubit Control”. In: *Proceedings of the IEEE Electron Devices Technology and Manufacturing Conference (EDTM). Submitted*. IEEE, 2026

12. E. Hua, S. van Ommen, K. Y. Yu, J. van Leeuwen, R. Bishnoi, H. Abunahla, S. Nur, S. Feld en R. Ishihara. “Memristor-Accelerated Neural Network Architectures for Energy-Efficient Quantum State Tomography”. In: *Proceedings of the IEEE Electron Devices Technology and Manufacturing Conference (EDTM). Submitted*. IEEE, 2026

11. E. Hua, S. Heemskerk, W. Braganc, H. Xun, A. Kaichouhi en R. Ishihara. “Ultra-Low-Noise Cryogenic Multilevel Memristors for Sub-100 $\mu$V Spin Qubit Biasing”. In: *Proceedings of the IEEE International Electron Devices Meeting (IEDM). Accepted*. San Francisco, USA: IEEE, 2025

10. E. Hua, T. Spyrou, M. Ahmadi, A. M. Syed, H. Xun, L. Braic, E. van der Veer, N. Elatab, A. Gebregiorgis, G. Gaydadjiev, B. Noheda, S. Hamdioui, R. Ishihara en H. Abunahla. “PdNeuRAM: Forming-Free, Multi-Bit Pd/$HfO_2$ ReRAM for Energy-Efficient Computing”. Nature Communications Engineering, in revision. 2025

9. E. Hua, H. Abunahla, G. Gaydadjiev, S. Hamdioui en R. Ishihara. “Multi-level forming-free HfO 2-based ReRAM for energy-efficient computing”. In: *2024 Device Research Conference (DRC)*. IEEE. 2024, p. 1–2

8. E. Hua, H. Abunahla en R. Ishihara. “Resistive Random Access Memory for performing forming-free resistive switching and method of manufacturing the memory”. Patent NL2036647B1. T. Delft. (Granted). 2025. URL: https://patents.google.com/patent/NL2036647B1

7. E. Hua, S. van Ommen, K. Y. Yu, J. van Leeuven, R. Bishnoi, H. Abunahla, S. Nur, S. Feld en R. Ishihara. “Neural Network Architectures for Scalable Quantum State Tomography: Benchmarking and Memristor-Based Acceleration”. In: *arXiv preprint arXiv:2507.23007* (2025)

6. E. Hua, T. Spyrou, M. Ahmadi, A. M. Syed, H. Xun, L. Braic, E. van der Veer, N. Elatab, A. Gebregiorgis, G. Gaydadjiev *e.a.* “PdNeuRAM: Forming-Free, Multi-Bit Pd/HfO2 ReRAM for Energy-Efficient Computing”. In: *arXiv preprint arXiv:2505.22789* (2025)

5. E. Hua, C. Glessi, C. W. Hagen en R. Ishihara. “"Novel Approach to a-Si Formation Using Eletron Beam Induced Processing with Liquid Heavy Polysilanes"”. Manuscript in submission to APL Materials. 2025

4. L. Landsmeer, E. Hua, H. Abunahla, A. Siddiqi, R. Ishihara, C. I. De Zeeuw, S. Hamdioui en C. Strydis. “Efficient Implementation of the Hodgkin-Huxley Potassium Channel via a Single Volatile Memristor”. In: *Frontiers in Neuroscience* 19 (2025), p. 1569397

3. J. van der Leeuven, E. Hua, R. Bishnoi, H. Abunahla, S. Nur en R. Ishihara. “Quantum State Tomography by leveraging a fully Spiking Variational Autoencoder on Neuromorphic Hardware”. *In preparation.* 2025

2. H. Xun, M. Fieback, S. Yuan, C. Wang, E. Hua, L. B. Poehls, H. Aziza, R. Cantoro, B. Rajendra, M. Taouil en S. Hamdioui. “In-Field Monitoring and Preventing Read Disturb Faults in RRAMs”. In: *IEEE Eur. Test Symp.* 2025, p. 1–6

1. H. Xun, M. Fieback, M. A. Yaldagard, S. Yuan, E. Hua, H. Aziza, M. Taouil en S. Hamdioui. “Robust Design-for-Testability Scheme for Conventional and Unique Defects in RRAMs”. In: *IEEE Int. Test Conf.)* 2024, p. 374–383. DOI: 10.1109/ITC51657.2024.00059

Publications non-related to PhD research:

7. E. Hua, S. Choi, S. Ren, S. Kim, G. Ali, S. J. Kim, W.-S. Jang, S. Joo, J. Zhang, S. Ji *e.a.* “Negatively charged platinum nanoparticles on dititanium oxide electride for ultra-durable electrocatalytic oxygen reduction”. In: *Energy & Environmental Science* 16.10 (2023), p. 4464–4473

6. E. Hua, S. Jin, X. Wang, S. Ni, G. Liu en X. Xu. “Ultrathin 2D type-II pn heterojunctions La2Ti2O7/In2S3 with efficient charge separations and photocatalytic hydrogen evolution under visible light illumination”. In: *Applied Catalysis B: Environmental* 245 (2019), p. 733–742

5. E. Hua, S. Jin, S. Ni en X. Xu. “Double perovskite compounds A 2 CuWO 6 (A= Sr and Ba) with p-type semiconductivity for photocatalytic water oxidation under visible light illumination”. In: *Inorganic Chemistry Frontiers* 6.8 (2019), p. 2096–2103

4. E. Hua, G. Liu, G. Zhang en X. Xu. “In situ fabrication of two-dimensional gC 3 N 4/Ba 5 Ta 4 O 15 nanosheet heterostructures with efficient charge separations and photocatalytic hydrogen evolution under visible light illumination”. In: *Dalton Transactions* 47.12 (2018), p. 4360–4367

3. X. Xu en E. Hua. “A kind of preparation method of photocatalyst preparation”. Chinese Patent CN109289883B. T. University. Granted. 11 sep 2018–11 mei 2021. URL: https://patents.google.com/patent/CN109289883B/en

2. T. Ma, R. Cao, X. Bao, E. Hua, M. Yang en Y. Wang. “Selective adsorption of trace H2O over O2 on Pt/Fe/Pt (111) surface of Pt-Fe catalyst”. In: *Applied Surface Science* 476 (2019), p. 387–390

1. S. Zhou, R. Meylaerts, E. Hua en L. Zhang. “Trust to test translation practices: A case study of Shanghai, China”. In: *Translation and Translanguaging in Multilingual Contexts* 10.2 (2024), p. 117–138



ISBN 978-94-6473-962-6