A Provenance-First, AI-Ready Data Platform for Large-Scale Memristor Experiments

Lai Gan [1] **Guoyang Huang** [1] **Deepika Yadav** [1] **Ben D. Rowlinson** [1] **Themis Prodromakis** [1]
[1]*Centre for Electronics Frontiers, School of Engineering, University of Edinburgh, Edinburgh, EH9 3BF, UK*. Correspondence to: Lai Gan L.Gan-9@sms.ed.ac.uk.

1. Introduction

Memristors are two-terminal electronic devices characterized by a pinched hysteresis loop in the current-voltage plane, as well as time-dependence and stochasticity [1]. They typically consist of a vertical stack of metal-insulator-metal (MIM), with a wide range of material combinations explored in the past 20 years [2, 3]. Memristors have been emerging with broad applications including neuromorphic systems,[4], multibit memory,[5], bio-processors,[6], and circuit calibration/tuning [7].

The advancement of memristive technology is inherently multidisciplinary, spanning materials, fabrication, characterization, compact modeling, and circuit/system integration; reliable and reproducible analysis therefore requires end-to-end provenance and traceable processing. We develop **STARS** (Semiconductor Traceable & AI-supported Research System) with memristors as the first target, unifying raw measurements, processing provenance, and derived features under versioned configurations so every reported result is traceable to its originating experiment. This auditability supports both conventional analysis and AI-assisted workflows—where verifying the basis of a conclusion matters as much as producing it—and, combined with automated acquisition and quality checks, enables unattended runs with analysis in parallel for faster iteration.

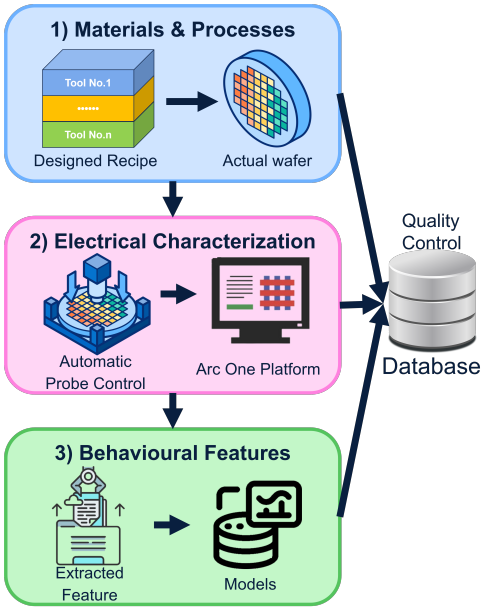

Fig. 1: STARS automated data collection pipeline. The workflow links (1) materials/process context, (2) device location & electrical characterization, and (3) behavioural feature generation, with quality check gates at each stage to preserve provenance for traceable, analysis-ready datasets.

2. Platform overview

2.1 Data acquisition and processing workflow

As shown in Fig. 1, STARS processes data in a staged workflow with quality control checks, so derived results can be linked back to the underlying measurements. The workflow consists of:

1. Materials and wafer context: record the fabrication recipe/material stack and the device location (wafer/die/device IDs and coordinates), with quality control checks for missing fields, inconsistent naming, and duplicate or conflicting identifiers.

2. Electrical characterization: import raw electrical traces and measurement settings from the ArC ONE platform [8], with quality control checks for missing or corrupted experiments, unrealistic signal values, and mismatches between the recorded protocol and the executed run.

3. Behavioural data products: extract derived features and model parameters, and perform quality control checks to check for out-of-range values, excessive missing data, and valid links to source measurements and processing configurations.

2.2 Schema design and provenance

As shown in Fig. 2, STARS links fabrication context, device location, measurements, and derived data products within a single relational database. Process information is stored at the recipe/layer level (including tool- and stack-related descriptors) and connected to individual devices through the wafer hierarchy (wafer–die–subdie–device). Each measurement run is recorded as an experiment, with parameters including applied voltage range, pulse duration, and other settings stored alongside it, while the high-volume point traces (e.g., voltage/resistance sequences and stimuli) are kept in a separate point-level table.

Derived behavioural outputs are stored in feature tables, where each row links back to its source experiment and references a configuration identifier that records the method and parameter settings used. To minimize duplicated information (e.g., repeatedly storing the same wafer/device information and parameter settings across many experiments), STARS stores shared metadata once and references it via identifiers.

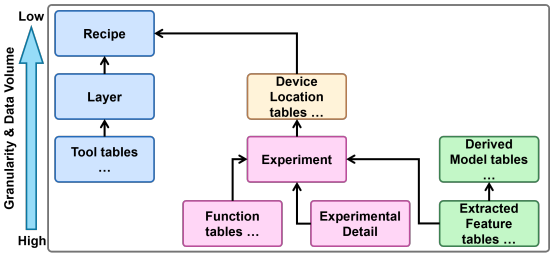

Fig. 2: STARS database schema overview. Fabrication context, device location, experiments, and measurements are linked to derived features and models through shared identifiers, supporting traceable queries and dataset export.

This layout follows standard database normalization practice (third normal form), [9] which removes redundancy and therefore reduces overall storage footprint, while keeping frequently used tables compact. As a result, routine filtering typically operate on device-, experiment- and feature-level tables first, and point-level traces are loaded only for the selected experiments when raw signals are needed; since each experiment may contain thousands of data points, avoiding point-table scans during initial selection markedly reduces query cost.

3. Demonstration and benchmarks

Table 1 summarises the quality-checked dataset and representative access costs. The current corpus spans 3 wafers, 75 dies, 6k devices, 170k experiments, and 160M point-level records, with five derived feature families; the framework scales to substantially larger volumes. For context, we benchmarked the partial hierarchical schema against flat-file baselines (e.g., CSV-style layouts) to quantify the cost of realistic joins and provenance-preserving filters; Table 1 reports the indexed SQLite results used in STARS.

Table 1: Quality-checked dataset scale and representative access costs for the indexed SQLite implementation used in STARS (hierarchical joins with provenance-preserving filters and aggregation over 1–20M point records).

Metric	Value
quality-checked dataset	
Wafers / Dies / Devices	3 / 75 / 6k
Experiments / Points	170k / 160M
Derived feature families	5
DB footprint	24 GiB
Access cost for indexed SQLite (benchmark)	
Group query time	0.46-1.05 s
DB footprint	0.12-1.09 GiB

4. Future AI workflows enabled by the STARS platform

STARS is designed to enable the following ongoing AI workflows. We outline the workflows here as future directions enabled by the platform.

(i) Autonomous measurement tuning: closed-loop selection of measurement parameters (e.g., forming windows, compliance settings, pulse schedules) under experimental noise and constraints, using data-efficient optimization such as (constrained or multi-objective) Bayesian optimization and active learning, as well as sequential decision-making formulations (e.g., contextual bandits) when exploration-exploitation trade-offs are critical[10, 11]. Candidate settings are proposed conditioned on device/material/location metadata and prior outcomes, and all actions and outcomes are logged back with full provenance for continual improvement.

(ii) Materials/location-to-behaviour learning: learning mappings from high-dimensional process/stack/geometry metadata and spatial factors to behavioural data products (performance-relevant features). The goal is to identify which fabrication and spatial factors most strongly explain observed behaviour and variability (often including material/process choices), by combining predictive modelling with model interpretation to separate explainable structure from residual stochasticity[12, 13].

(iii) Knowledge extraction from learned policies: when closed-loop policies succeed, they may capture latent regularities in the underlying device physics. These regularities can remain as deployable black-box decision models, or, where possible, be distilled into interpretable abstractions, for example, compact rule-based heuristics, probabilistic yield/variability models, or simplified analytic forms that summarize dominant physical trends [14, 15].

5. Conclusion

We present STARS, a data platform that organizes large-scale memristor experiments into a traceable evidence base. STARS links fabrication details, device location, measurements, and derived behavioral features within a single schema, and applies quality control checks throughout the workflow. It also supports automated data acquisition and structured dataset access, so researchers can quickly filter experiments, build consistent subsets, and export analysis-ready tables for statistics and AI. We report the current dataset scale and representative benchmark costs to demonstrate practical usability on open-source hardware. In general, STARS reduces manual overhead, improves comparability between experiments, and provides a firm foundation for closed-loop optimization and explainable AI workflows for accelerating memristor development.

Acknowledgments

All authors appreciate UK Research and Innovation (UKRI) and Engineering and Physical Sciences Research Council (EPSRC) funding for the AI Hub for Productive Research and Innovation in Electronics (APRIL) AI Hub [grant number EP/Y029763/1].

References

[1] Leon Chua. Resistance switching memories are memristors. *Applied Physics A*, 102(4):765–783, 2011.

[2] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. The missing memristor found. *Nature*, 453(7191):80–3, 2008.

[3] Mostafa Shooshtari, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. Review of memristors for in-memory computing and spiking neural networks. *Advanced Intelligent Systems*, 11(11):e202500806, 2025.

[4] Alexander Serb, Johannes Bill, Ali Khiat, Radu Berdan, Robert Legenstein, and Themis Prodromakis. Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses. *Nature communications*, 7(1):1–9, 2016.

[5] Spyros Stathopoulos, Ali Khiat, Maria Trapatseli, Simone Cortese, Alexantrou Serb, I. I. Valov, and Themistoklis Prodromakis. Multibit memory operation of metal-oxide bi-layer memristors. *Scientific Reports*, 7, 2017.

[6] Grahame Reynolds, Xiongfei Jiang, Alexander Serb, Themis Prodromakis, and Shiwei Wang. An integrated cmos/memristor bio-processor for reconfigurable neural signal processing. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5, 2023.

[7] Zhaoguang Si, Chaohan Wang, Xiongfei Jiang, Zheyi Li, Guoyang Huang, Alexander Serb, Themis Prodromakis, Shiwei Wang, and Christos Papavassiliou. Memristor-assisted background calibration for sar adcs: A feasibility study. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(9):3497–3508, 2023.

[8] ArC Instruments Ltd. ArC ONE User's Guide. PDF manual. Accessed: 2026-02-06.

[9] C. J. Date. *An Introduction to Database Systems*. Addison-Wesley, Boston, 8 edition, 2003.

[10] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599, 2010.

[11] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, 2012.

[12] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[13] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[14] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[15] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.