Cosmo Nazzareno Santoni
Email: santonicosmo@gmail.com | cosmo.santoni@imperial.ac.uk | Tel: +44 7444 403542
Links: GitHub | Google Scholar | Personal Website | LinkedIn
Location: London, UK (Available to relocate)

Machine learning researcher, engineer and PhD candidate in Applied Mathematics at Imperial College London specialising in sequence models, state-space architectures, and simulation-based inference for decision-making and large language models (LLMs), LLM robustness and safety, with 125,000× speed-ups over traditional simulators while maintaining calibrated uncertainty estimates. Deployed in production for World Health Organisation (WHO) malaria programme planning and UK government crisis response; publications include Nature and The Lancet with work presented at a NeurIPS 2025 workshop and manuscripts under review at top-tier ML conferences 2026 and TMLR. Creator of mamba2-jax, a JAX/Flax Mamba-2 implementation merged by the Bonsai JAX team into Google's official jax-ml/bonsai models repository.

SELECTED IMPACT

- Contributed the **Mamba-2** (JAX/NNX) implementation to **Google's jax-ml/bonsai model zoo**, including pretrained-weight loading and state-space caching (up to **35×** inference speedup). Ongoing contributor; co-authoring a Google Open Source Blog post with the JAX Bonsai team.
- Developed **state-space and RNN neural surrogate architectures** for large-scale agent-based models, achieving **~125,000× speedups** against reference agent-based simulator, deployed **in production for WHO** malaria intervention planning and international government decision support.
- Received **SAGE Award** from **Sir Patrick Vallance & Professor Chris Whitty, UK Government Chief Scientific & Medical Officers** for modelling and data support enabling evidence-based COVID-19 policy.

PROFESSIONAL EXPERIENCE

Full-Stack Machine Learning Research Software Engineer – *Imperial College London, U.K., 2023 – Present*

- Designed a two-stage neural emulator that replaces a 24-minute stochastic malaria simulator with 10.5ms inference (125,000x speedup), achieving sub-1% error across 464 validation scenarios spanning 8 intervention types. First author on TMLR submission.
- Built the end-to-end ML platform from synthetic data generation over 6.89B data points (DuckDB, sub-second analytical queries) through PyTorch training with hot-swappable sequence architectures (LSTM, GRU, Mamba-2), CUDA mixed-precision training, and Optuna hyperparameter search to deterministic model serving APIs.
- Shipped five open-source Python packages covering inverse calibration, training, inference, and deployment. Adopted by research teams across Imperial and partner institutions, reducing model development cycles from months to days. Presented at international workshops and used in active deployment programmes.
- Partnered with WHO to deploy the emulator for national malaria intervention planning, translating millisecond-latency model queries into policy-ready projections where hours of HPC compute were previously required.

Machine Learning Researcher – *University of Cambridge, Department of Computer Science, UK., & German Centre for Artificial Intelligence (DFKI), DE., 2022 – 2023*

- Engineered neural ODE architectures with hard physical constraints for complex dynamical systems. Implemented physics-informed loss functions and automatic differentiation in PyTorch / Julia, achieving strong performance on irregular sparse time-series while maintaining clinical interpretability.
- Designed training and deployment infrastructure for multi-scale dynamical systems, building pipelines supporting multiple optimisation algorithms (Adam, Levenberg-Marquardt), Bayesian uncertainty quantification, and sensitivity analysis suitable for safety-critical settings.
- Drove architecture decisions for Neural Universal Differential Equations, balancing expressiveness vs interpretability. Developed 1D/2D neural ODE architectures that enabled previously intractable dynamical systems where unconstrained data-driven baselines failed physical consistency checks.

Research Assistant – *COVID-19 Real-time Modelling, Imperial College London, U.K., 2021 – 2023*

- Owned real-time inference pipeline for national-scale epidemic time-series, delivering <24hr from raw data to policy recommendations. Outputs directly informed UK lockdown policy during the national emergency.
- Maintained open-source inference libraries under active use by UK SAGE during the pandemic. Managed breaking changes, backward compatibility, and emergency bug fixes while ensuring reproducibility across distributed teams.
- Engineered automated reporting infrastructure generating statistical analyses, visualisations, and forecasts for systems running continuously for 18+ months, supplying weekly briefings to UK Chief Scientific Advisors.

TECHNICAL SKILLS & LANGUAGES

- ***Languages**: Python, R, SQL, Bash (Linux)*
- ***ML / DL:** JAX, PyTorch, JAX-CFD, **Transformers**, Mamba2, RNNs/LSTMs, XGBoost, Optuna, NumPy, Pandas*
- ***Data & Storage:** DuckDB, HDF5*
- ***HPC & Acceleration**: CUDA, XLA, MPS, Automatic Mixed Precision (AMP); CPU/GPU clusters; Google Cloud TPUs / GCP;*
- ***Tooling:** Git / GitHub, CI/CD (GitHub Actions), Linux, Docker (deployment exposure)*
- ***Research focus:** RNNs, state-space models (Mamba-2), Transformers and large language models (LLMs), generative modelling (flow matching), neural ODEs, simulation-based inference (SBI), Bayesian optimisation and modelling, continual learning, conformal prediction, calibration and uncertainty quantification (UQ).*

EDUCATION

- **PhD in Applied Mathematics,** *Imperial College London, U.K., January 2025 – December 2027*
  Thesis: "Towards State Space and Continual Learning Models for Large-Scale Agent-Based Simulators"
- **MSc. Epidemiology (Merit),** *Imperial College London, U.K., 2020 – 2021*
- **BSc. (Hons.) Mathematics with Economics,** *Aston University, U.K., 2015 – 2019*

SYSTEMS, OPEN SOURCE & RESEARCH SOFTWARE

Production ML Systems & Flagship Libraries

- *mamba2-jax: JAX/Flax Mamba-2 implementation. Core architecture (PR #103) with state-space caching (PR #131), merged into Google's jax-ml/bonsai; ongoing collaboration with the Bonsai team.*
- *MINTverse: production orchestration (R/Python) delivering neural emulation for WHO operational decision-support.*
- *hiddenstate.io: ML market intelligence engine aggregating daily across arXiv, GitHub, Hugging Face, OpenReview and other independent sources, clustering and scoring via a multi-signal W-index. 1,000+ daily visitors. Featured in Claude Builder Club @ Imperial (Anthropic-affiliated program).*

Open-source libraries & tooling

- *claude-code-cmv: Context memory virtualisation for Claude Code. DAG-based snapshot, branch, and structurally lossless trim primitives for LLM session state; Technical report: arXiv:2602.22402. 60+ stars and 7 forks.*
- *mamba2-triton-guard: Lightweight patcher that stubs Triton and guards version checks so mamba2_torch imports cleanly on macOS M1–M5 (CPU / MPS), enabling SSM experimentation without GPU triton-only constraints.*
- *Epireview: production data extraction and automated figures / tables for WHO PERG; used by international teams*
- *Sircovid, Spimalot, MCState: UK COVID-19 Bayesian toolkit (adaptive PMCMC, inference, forecasting).*

PUBLICATIONS & RESEARCH OUTPUT

Selected Publications:

- *Perez-Guzman, P. N., Knock, E., et al. "Epidemiological drivers of transmissibility and severity of SARS-CoV-2 in England." **Nature Communications (co-author)***
- *Imai, N., Rawson, T., et al. "Quantifying the impact of delaying the second COVID-19 vaccine dose in England: a mathematical modelling study" **The Lancet Public Health (co-author)***
- *Multiple publications under PERG including **Lancet Infectious Disease, Lancet Global Health & Lancet Microbe***

Manuscripts under Preparation/Review:

- *Charles, G., Santoni, N. C., et al. "Tokenised Flow Matching for Hierarchical Simulation Based Inference"*, **Manuscript under review (venue withheld due to double-blind policy; co-author)**
- *Santoni, N. C., et al. "Emulating Stochastic Simulators with Latent Inputs via Inverse-Forward Decomposition" **TMLR**, **2026**. **(target venue; first-author)***
- *Santoni, N. C., et al. "Compiler-First State Space Duality and O(1) Autoregressive Caching for Inference" **TMLR**, **2026**. **(target venue; first-author)***
- *Santoni, N.C., et al. "Linear-Time Amortized Simulation-Based Inference for Spatial Models via Selective State Spaces" **NeurIPS, 2026 (target venue; first-author)***
- *Santoni, N. C., Schwarz, J. R., et al. "Freeze-Thaw Bayesian Optimisation for Accelerated Data Mixture Learning", (**In preparation; first-author)***

Conferences, Workshops & Presentations:

- **Tokenised Flow Matching for Hierarchical Simulation-Based Inference**, *NeurIPS 2025 Workshop on Frontiers in Probabilistic Inference: Sampling Meets Learning*, *San Diego, USA*

ACADEMIC SERVICE & VOLUNTARY WORK

- **Foundational Machine Learning Research Working Group Co-chair**, Imperial College London, U.K., 2026 –
- **Curator, Amphibian & Malaria Collections**, *Museum of Life Sciences, King's College London*, U.K., 2025 – 2026
- **Departmental Seminar Series Co-Organiser**, *Imperial College London, U.K.*, 2023 – 2025