Cosmo Nazzareno Santoni
Email: santonicosmo@gmail.com | cosmo.santoni@imperial.ac.uk | Tel: +44 7444 403542
Links: GitHub | Google Scholar | Personal Website | LinkedIn

ML researcher at the intersection of neural surrogates and real-time decision support, bridging theoretical modelling and practical deployment. I develop principled sequence-model architectures (neural ODEs, state-space models) achieving 10^5× speed-ups over traditional simulators whilst maintaining calibrated uncertainty estimates. Deployed in production for WHO malaria programme planning and UK government crisis response; publications include *Nature* and *The Lancet*, with work under review at *ICML/KDD*.

SELECTED IMPACT

- Neural surrogate methods achieving 125,000× speedups deployed in production for WHO malaria intervention planning and international government decision support.
- Published in **Nature** and **The Lancet**; work under review at **ICML** and **KDD**.
- **SAGE Award** from UK Government **Chief Scientific & Medical Officers** for COVID-19 response.

PROFESSIONAL EXPERIENCE

Machine Learning Researcher (Technical Analyst) – *Imperial College London, U.K., 2023 – Present*

- Led end-to-end design and research of ML-based simulation infrastructure, replacing 24-minute HPC agent-based computations with sub 10.5ms neural inference (over 125,000x speedup, 99.8% fidelity). Enabled real-time what-if analysis previously impossible at decision-making speed deployed for government policy across multiple countries.
- For typical funded annual modelling requests (~1M scenarios), cut energy from ~22–43 MWh on consumer/HPC CPUs (≈£5.9k–£11.3k) to ~0.13 kWh with 10.5ms neural inference.
- Architected MINTverse production ML platform processing 100+ scenarios per second on consumer GPU. Designed modular neural multi-task learning emulation architecture with hot-swappable RNN/Mamba-2 heads for sequence modelling (PyTorch + CUDA + AMP), built data infrastructure (DuckDB handling 574M records with sub-second queries), implemented model serving APIs, and deployed with deterministic versioning. System enabled policy decisions at millisecond latency where hours were previously required.
- Engineered high-throughput analytics layer processing 6.89B data points with sub-second query performance.
- Built open-source modular ML infrastructure (MINTverse: segMINT, estiMINT, MINTe, MINTer, and more) reducing internal teams' research iteration cycles from months to days.
- Partnered cross-functionally with WHO to translate neural emulation outputs into deployment guidance. Bridged technical performance and policy requirements for international malaria intervention programmes.

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

- Engineered neural ODE framework for complex biological systems (tumour growth, epidemic dynamics) with hard physical constraints. Implemented physics-informed loss functions and automatic differentiation in PyTorch/Julia, achieving high predictive accuracy on irregular sparse time-series while maintaining clinical interpretability.
- Designed production infrastructure handling multi-scale dynamical systems for tumour growth modelling. Built training pipelines supporting multiple optimisation algorithms (Adam, Levenberg-Marquardt), implemented Bayesian uncertainty quantification, and integrated sensitivity analysis meeting clinical safety requirements.
- Drove architecture decisions for Neural Universal Differential Equations balancing expressiveness vs interpretability. Developed 1D/2D neural ODE architectures enabling previously intractable biological systems where pure data-driven approaches failed physical consistency checks.

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

- Owned real-time inference pipeline delivering daily briefings to UK government under extreme time pressure (<24hr from raw data to policy recommendations). Particle MCMC implementation in C++/R processed national-scale epidemic data with sub-hour turnaround when decisions affected millions. Zero tolerance for errors—every output directly informed UK lockdown policy during national emergency.
- Maintained mission-critical modelling software (sircovid, spimalot, MCState) under active use by UK SAGE during pandemic. Managed breaking changes, backward compatibility, and emergency bug fixes while ensuring reproducibility across distributed teams working under crisis conditions. Software became foundation for international epidemic response efforts.
- Engineered automated reporting infrastructure generating statistical analyses, visualisations, and forecasts with <24hr latency end-to-end. System ran continuously for 18+ months supplying weekly briefings to UK Chief Scientific Advisors with zero missed deadlines.
- Received SAGE Award for Modelling and Data Support from UK Government Chief Scientific and Medical Officers recognising technical contributions enabling evidence-based crisis response at national scale.

TECHNICAL SKILLS & LANGUAGES

- **Languages:** Python, R, Julia, C++ (familiar), C#, SQL, Bash (Linux)
- ***ML / DL Systems:*** *PyTorch, JAX, JAX-CFD, MLX, XGBoost, Optuna, NumPy, Pandas*
- ***Data & Storage:*** *DuckDB, HDF5*
- ***HPC & Acceleration:*** *CUDA, MPS, Automatic Mixed Precision (AMP); CPU/GPU clusters*
- ***Tooling:*** *Git/GitHub, CI/CD (GitHub Actions), Linux*

EDUCATION

- **Full-time PhD in Applied Mathematics,** *Imperial College London, U.K., January 2025 – December 2027*
  Thesis: "Accelerating Real-Time Agent-Based Simulations with State-Space Neural Surrogates and Graph Neural Networks"
- **MSc. Epidemiology (Merit),** *Imperial College London, U.K., 2020 – 2021*
- **BSc. (Hons.) Mathematics with Economics (2:1),** *Aston University, U.K., 2015 – 2019*

PUBLICATIONS & RESEARCH OUTPUT

Production Systems (World Health Organisation)

- *MINTverse: production orchestration (R/Python) delivering neural emulation for WHO operational decision-support.*
- *Epireview: production data extraction/figures/tables service for WHO PERG; used by international review teams.*
- *RADAR (in-dev): state-space machine error calibration for time-series emulators (private; results on request)*

LLM, alignment & safety

- *mamba2-triton-guard: tiny patcher that stubs Triton and guards version checks so mamba2_torch imports on macOS M1–M5 (CPU/MPS)*
- *mamba2_torch contribution (PR under review): Added import guard for MPS, gated fast-path, and fixed padding edge case; improves stability on M-series and avoids silent shape errors.*
- *Metacognitive / Continuity Prompting of Claude - independent technical report reviewed by Anthropic Trust & Safety: systematic behavioural analysis identifying non-uniform helpfulness patterns with proposed alignment evaluation framework*
- *ALETHIA + KAIROS: metacognitive control and selective acceptance for LMs, with full audit artefacts.*

Additional Open-source libraries

- *Sircovid, Spimalot, MCState: UK COVID-19 Bayesian toolkit (adaptive PMCMC, inference, forecasting).*

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)***
- *Further Publications Under PERG including Lancet Infectious Disease, Lancet Global Health & Lancet Microbe*

Manuscripts under Preparation/Review:

- *Charles, G, et al. "Tokenised Flow Matching for Hierarchical Simulation Based Inference"* **International Conference on Machine Learning,** *under review, 2026* ***(co-author)***
- *Santoni, N. C., et al. "Deploying State-Space Neural Emulators for Real-Time Malaria Policy Decision Support"* ***ACM SIGKDD International Conference on Knowledge Discovery and Data Mining****, in preparation, 2026.*

CONFERENCES & PRESENTATIONS

- **Epireview: Hands-on Workshop for Public Health & Epidemiology Researchers,** *Infectious Disease Modelling Conference, Bangkok, Thailand, November 2024*
- **Applying Neural Network Emulation to Assess the Impact of Pyrethroid-Pyrrole Bed Nets on Malaria in Africa,** *9th International Conference on Infectious Disease Dynamics, Bologna, Italy, November 2023*
- **Investigating Parameterisation and Inference Trade-Offs in Stochastic and Deterministic Epidemic Models**, *9th International Conference on Infectious Disease Dynamics, Bologna, Italy, November 2023*

ACADEMIC SERVICE & VOLUNTARY WORK

- **Curator, Amphibian & Malaria Collections**, *Museum of Life Sciences, King's College London*, U.K., 2025 –
- **Departmental MRC GIDA Seminar Series Co-Organiser**, *Imperial College London, U.K.*, 2023 –
- **Lay Grant Reviewer**, *University College London & Parkinson's UK*, U.K., 2019 – 2022