Sara
Molas Medina, PhD

ML Researcher with interdisciplinary training in Systems Neuroscience and Mechanistic Interpretability.
Experienced in training and mechanistically analysing deep learning models.
Strong background in statistical modeling, experimental design, and high-dimensional data analysis.

saramolas18@gmail.com saramolasmedina SaraMolas saramolas.github.io

Professional Experience

Independent Researcher, funded by Coefficient Giving, Apr 2026 – Present
Currently investigating the internal representations underlying compositional generalization in diffusion models (paper in progress), while developing a broader research agenda in AI safety.

Research Fellow, Athena programme, Nov 2025 – Feb 2026
Trained conditional diffusion models to study compositional generalization emergence and analysed training dynamics based on Singular Learning Theory.
Mentored by Jesse Hoogland (Timaeus) - publication in progress.

Independent Researcher, July – September 2025
Applied Sparse Autoencoders to synthetic biological neural data to extract latent features (NeurIPS workshop paper).

Data Scientist, VodafoneThree, Mar 2025 – Mar 2026
Built large-scale ML pipelines improving campaign performance by 30%.
Developed an LLM-based agent (LangChain) to interact with customers and recommend products.
Applied LLM-based analysis of customer calls to identify revenue increase opportunities.

Research Fellow, SPAR, Feb - June 2025
Conducted mechanistic interpretability experiments in neural networks.
Performed targeted ablation studies and activations visualization to study feature superposition.
Mentored by Stefan Heimersheim (FARAI) – NeurIPS workshop paper.

PhD Researcher, *University College London, Sep 2019 – Dec 2024*
Led independent research on internal representations in biological neural networks.
Developed end-to-end analysis and modelling pipelines for high-dimensional neural population data.
Applied Bayesian decoders, dimensionality reduction, and statistical modelling.

Education

PhD Systems and Computational Neuroscience, UCL & QMUL (UK), 2019 – 2024
Funded by LIDo Doctoral Training programme (5% acceptance rate)

MSc Neuroscience (Distinction), UCL (UK), 2018 - 2019

BSc Biomedical Sciences (First Class Honours), UAB (Spain), 2014 – 2018

Selected Publications in Mechanistic Interpretability and Deep Learning *(* denotes equal contribution)*

- **Molas-Medina**. Interpreting the representations underlying compositional generalization in a diffusion model (In prep).
- Bhagat*, **Molas-Medina***, Giglemiani, Heimersheim. Compressed Computation is (probably) not Computation in Superposition. *NeurIPS, Mechanistic Interpretability workshop paper,* 2025.
- Bhagat, Pouget, **Molas-Medina**. A pipeline for interpretable neural latent discovery. *NeurIPS, Data on the Brain & Mind findings workshop paper, 2025.*

Additional Research Activities

- Open Philanthropy technical AI safety RFP Research Grant (Final Round, 2025)
- Peer reviewer: NeurIPS 2025 workshop UniReps: Unifying Representations in Neural Models
- ARENA (AI alignment Research Engineer Accelerator) course (2025)
- Machine Learning Summer School (Stellenbosch University, South Africa, 2023)

Machine Learning & AI
Mechanistic interpretability
Deep learning
Supervised learning
Unsupervised learning

Engineering
Python (incl. PyTorch, sklearn, numpy, pandas)
Data/ML pipelines
Version control (Git)

Data analysis
Statistics
Data processing
Data exploration
Predictive modeling
Experimental design

Additional skills
Independent research
Group research
Cross-disciplinary
Scientific writing