FABIO VALERIO MASSOLI

AI/ML RESEARCHER

CONTACT

+31 6 58941247

fabiovaleriomassoli@gmail.com

519, 1087LG, Amsterdam (NL)

LinkedIn
Scholar

SKILLS

- **AI frameworks:** PyTorch, Lightning, Hugging Face ecosystem (transformers, TRL, accelerate, PEFT, datasets, lighteval), vLLM, verl
- **Distributed training:** FSDP, DeepSpeed
- **Programming:** Python, C++, bash
- **Simulation:** Geant4, FairRoot
- **Other:** git, NumPy, PyMC, scikit-learn, pandas, Matplotlib, seaborn, PIL, OpenCV, h5py

LANGUAGES

- **Italian** (Native)
- **English** (Fluent - C1)
- **Dutch** (Basic - A1)

AWARDS

- **Student Mobility Grant. PhD, (€12,000)** - visiting researcher at Columbia University, NY. XENON collaboration.
- **CNR Young Mobility Grant (€15,000) -** visiting researcher at USC. Project focused on adversarial attack detection
- **Nvidia Research Grant (won Nvidia Titan-X)** for research proposal on brain modelling

SUMMARY

Research scientist (PhD in Physics) with 7+ years experience advancing foundation models, efficient reasoning, looped language models, and RL post-training for LLMs. Introduced the Conditional Information Bottleneck framework for budget forcing (NeurIPS 2026) and a looped-transformer design with 4x KV-cache reduction on Ouro models. Author at NeurIPS, ICML, and other venues; co-inventor on 10+ US patents and applications; #1 globally in the ITU DeepSense 6G Challenge. Hands-on in Python and large-scale distributed training and inference (FSDP, DeepSpeed, vLLM, verl). Lead a team of 3 researchers.

WORK EXPERIENCE

Qualcomm AI Research, Amsterdam Sep 2021 - Present

AI/ML Researcher

- Designed the **Conditional Information Bottleneck** framework for budget forcing (**NeurIPS 2026**), achieving Pareto-optimal efficiency across domains and model families.
- Engineered a **looped-LM** architecture **reducing KV footprint by 4x** (32GB to 8GB) on SOTA Ouro models, significantly lowering inference serving costs.
- Developed a **training-free optimization reducing FLOPs and KV memory by 25%** in **looped models** with zero performance drop, with a strong focus on production-ready model optimization.
- Orchestrated **distributed RL training** pipelines yielding up to **5x CoT length reduction**; core contributor to the on-device reasoning **NeurIPS 2025 demo**.
- **Lead a 3-person research team** shaping the strategic AI roadmap for RL in LLM post-training, reasoning, and efficient looped models.

ISTI, Pisa Mar 2018 - Aug 2021

AI/ML Researcher

- Designed and implemented state-of-the-art multi-resolution deep learning models for complex pattern recognition tasks.
- Developed novel algorithms for multi-resolution adversarial detection and state-of-the-art outlier detection.

Information Science Institute (ISI) - USC, LA Oct 2019 - Dec 2019

AI/ML Visiting Researcher

- Developed detection algorithms for adversarial attacks by analyzing latent space trajectories in deep learning models.

SysData s.p.a., Bologna Sep 2017 - Mar 2018

Android Software Engineer

- Led a team focused on UX optimization.
- Designed and developed UI for Android applications and libraries for reactive programming.

INFN, Bologna Jan 2015 - Aug 2017

Physics Researcher

- Led a quantitative research group modeling complex stochastic processes in high-dimensional spaces using advanced Monte Carlo methods.
- Designed, implemented, and validated large-scale Monte Carlo simulations for the XENON1T experiment.
- Developed robust, scalable codebases for advanced statistical modeling, including Bayesian inference techniques, applied to large, complex datasets.

REVIEWER ACTIVITY

Reviewer: ICML, NeurIPS, ICLR

PATENTS APPLICATIONS

Full list of patents and patent applications available on my scholar page: *scholar/fvmassoli*

F.V. Massoli, et al.: Reinforcement Learning for inverse problem simulation models; US Patent App. 18/661,511 [link]

F.V. Massoli, et al.: *Machine Learning for wireless channel model*; US Patent App. 18/158,885 [link]

F.V. Massoli, et al.: *Fusion models for beam prediction*; US Patent App. 18/340,671 [link]

G. Silvestri, **F.V. Massoli**, T. Orekondy, A . Behboodi, J.B. Soriaga: *Adaptive Acquisition for compressed sensing*; US Patent App. 18/475,995 [link]

F.V. Massoli, et al.: *Machine learning for wireless channel estimation*; US Patent App. 18/846,537 [link]

Columbia University, New York Aug 2013 - Dec 2013
Visiting Researcher

- Performed Monte Carlo simulations for the NeriX experiment.
- Applied statistical data analysis techniques to experimental data.
- Utilized Machine Learning techniques for signal analysis and pattern recognition in scintillation data.

CERN, Geneva Jun 2010 - Sep 2010
Visiting Researcher at Large Hadron Collider (LHC)

- Analyzed large-scale experimental data from the CMS experiment Drift Tubes.
- Developed and applied Monte Carlo simulations to study complex event formation.
- Designed algorithms for event tagging and noise reduction in high-volume data streams.

MOST RECENT PUBLICATIONS

Full list of publications available on my scholar page: *scholar/fvmassoli*

Accepted / Published

- **F.V. Massoli** et al.: *Reasoning as Compression: Unifying Budget Forcing via the Conditional Information Bottleneck*; [arXiv] **NeurIPS 2026**
- **F.V. Massoli** (core contributor) et al.: Efficient Reasoning on the Edge; [arXiv] **white paper + NeurIPS25 demo**
- A.H.C. Correia, **F.V. Massoli**, C. Louizos, A. Behboodi: An Information Theoretic Perspective on Conformal Prediction; [paper] **NeurIPS 2024**
- A. Kuzina, K. Pratik, **F.V. Massoli**, A. Behboodi: *Equivariant Priors for Compressed Sensing with Unknown Orientation;* [paper] **ICML 2022**

Under Review

- V.C. Vendrell, A.P. Masdemont, N. Grillo, J. Ros-Giralt. A. Behboodi, **F.V. Massoli**: *Memory-Efficient Looped Transformer: Decoupling Compute from Memory in Looped Language Models*; [arXiv] Under Review **ICLR 2027**
- Aayush Mishra, A.P. Masdemont, V.C. Vendrell, J. Ros-Giralt. A. Behboodi, **F.V. Massoli**: *Enabling Dynamic Computation in Looped LMs;* [arXiv] Under Review **ICLR 2027**
- **F.V. Massoli** et al.: *When do stored facts become available in looped transformers?*; [arXiv - coming soon] Under Review **ICLR 2027**

COMPETITIONS

First Place, ITU DeepSense 6G Beam Prediction Challenge 2023

- Designed and implemented a novel transformer-based architecture for a high-dimensional prediction task.
- Achieved #1 rank on the leaderboard for both in-distribution and out-of-distribution test scenarios, demonstrating the model's robustness against regime shifts.

EDUCATION

Ph.D. in Physics Jan 2012 - Dec 2015
University of Bologna and Columbia University (NY)

- Research focused on Monte Carlo simulations, Statistical Analysis, and Machine Learning applications in High Energy Physics (HEP).
  - Top 3% PhD student (Astrophysics department)

MS in Physics, University of Bologna Jan 2010 - Dec 2011

Bachelor in Physics, University of Bologna Oct 2006 - Dec 2009