

Curriculum vitae di *Simone Scardapane*

Sottosettori ERC primari (max 3): **PE6_11, PE6_7**

PERSONAL DETAILS

Family name, First name: Scardapane, Simone

Birthdate: 28/10/1988

Researcher unique identifier(s):

- ORCID: <https://orcid.org/0000-0003-0881-8344>
- Scopus: <https://www.scopus.com/authid/detail.uri?authorId=55772102700>
- Google Scholar: <https://scholar.google.com/citations?user=aSuosYoAAAAJ>

URL for web site: <https://www.sscardapane.it/>

• **Education and key qualifications**

05/2016 PhD in Information and Communication Technologies

Department of Information Engineering, Electronics and Telecommunications (DIET)

Sapienza University of Rome (Rome, Italy)

Supervisor: Prof. Aurelio Uncini

2011 Master's degree in Artificial Intelligence and Robotics (110 *magna cum laude*)

Sapienza University of Rome, Rome, Italy

• **Current position(s)**

2025 - Present Associate Professor

Department of Information Engineering, Electronics and Telecommunications (DIET)

Sapienza University of Rome (Rome, Italy)

• **Previous position(s)**

2022 - 2025 Tenure-track Assistant Professor (*Ricercatore a Tempo Determinato di Tipo B*)

Department of Information Engineering, Electronics and Telecommunications (DIET)

Sapienza University of Rome (Rome, Italy)

2019 - 2022 Assistant Professor (*Ricercatore a Tempo Determinato di Tipo A*)

Department of Information Engineering, Electronics and Telecommunications (DIET)

Sapienza University of Rome (Rome, Italy)

2016 - 2019 Post-doctoral fellow (*Assegnista di Ricerca*)

Department of Information Engineering, Electronics and Telecommunications (DIET)

Sapienza University of Rome (Rome, Italy)

RESEARCH ACHIEVEMENTS AND PEER RECOGNITION

Research achievements

Below I provide a selection of 8 recent research contributions, accompanied by a brief description of their scientific and technological impact. Broadly, my research focuses on the design, analysis, and deployment of novel neural network architectures, which represent the core computational paradigm underlying modern AI systems. A central goal of my work is to develop models that are not only accurate, but also efficient, adaptive, and interpretable, thereby enabling their use in large-scale, real-world technological systems.

On the methodological side, recent contributions have addressed key challenges in the field, including improving the efficiency of large-scale models [1, 3], extending deep learning to graph and topological domains [4, 7], and advancing methods for interpretability and model understanding [5]. On the applicative side, these techniques have been applied across a wide range of scientific and technological domains, including telecommunications [2], paleography and cultural heritage [5], high-energy physics [6], and remote sensing [8], often in close collaboration with domain experts.

As a reference over my entire career, I am the author of approximately 150 peer-reviewed publications (conferences and journals), with a current h-index (Scopus) of 31 and 5,026 citations at the time of writing.

[1] Efficient AI Architectures

One of the main limitations of current AI systems, including large language models, is the high computational cost required at inference time, which hinders their deployment in real-world settings. In recent work, my group has focused on the design of *dynamic neural architectures*, capable of adapting their computational footprint to the complexity of each input or to external constraints. This line of research includes the development of *adaptive computation modules* (AAAI, 2025), a *dynamic-k mixture-of-experts* component (NeurIPS, 2024), and adaptive models for structured data such as point clouds. Across multiple domains, including computer vision, speech processing, and language modeling, these approaches achieve reductions of 30–50% in computational cost at inference time while preserving predictive performance. More broadly, this work contributes to the development of scalable and resource-efficient AI systems, a key requirement for their integration into next-generation technological infrastructures.

[2] AI-Native Communication Systems

In recent work, particularly within the 6G-GOALS EU project (SNS-JU),¹ we have explored the integration of deep learning models directly into the communication pipeline, moving beyond the traditional separation between source coding, channel coding, and signal processing. In this context, we developed the concept of *AI-native communications*, where neural networks model the entire transmission process end-to-end. A central idea in this line of work is that of *token-based communication*, where tokens (analogous to those used in large language models) replace classical bits as the fundamental unit of information exchange. Building on this paradigm, we proposed several techniques to dynamically control the quantity and semantic relevance of transmitted information, leveraging adaptive models as described in [1]. These contributions have been validated across multiple venues (including IEEE Transactions on Machine Learning in Communications and Networking, Neurocomputing, Asilomar, and IEEE GLOBECOM). This line of research contributes to the development of next-generation communication systems that are inherently adaptive, data-driven, and tightly integrated with AI models, opening new directions for efficient and semantic communication in future network infrastructures.

[3] Efficient and Reliable Large Language Models

Large language models (LLMs) have become a central component of modern AI systems, but their deployment at scale raises important challenges in terms of efficiency, memory usage, and reliability. In this context, my recent work has focused on both understanding the internal properties of these models and developing methods to improve their practical performance. On the efficiency side, we proposed novel strategies for compressing key components of LLM inference, such as the *KV cache*, significantly reducing memory requirements while preserving accuracy. These methods have been adopted in open-source systems, including NVIDIA's KVPress library. Complementing this, we investigated fundamental properties of modern LLMs, including activation sparsity, representation alignment, and the emergence of phenomena such as attention sinks in diffusion-based

¹ <https://sites.google.com/view/6ggoals/home>

language models (ACL, 2026; NeurIPS, 2026; CVPR, 2025). Together, these contributions provide both practical tools and theoretical insights for improving the scalability, robustness, and interpretability of large-scale AI systems.

[4] Topological and Geometric Deep Learning

Standard deep learning models rely on fixed and relatively simple data structures, such as grids or sequences, which are often inadequate to capture complex relational patterns present in real-world data. In this context, my work has contributed to the emerging field of *topological and geometric deep learning*, which aims to extend neural networks to more expressive domains such as graphs, hypergraphs, and higher-order structures. Recent contributions include the development of benchmarking tools for topological learning (TopoBench), the study of fundamental challenges such as heterophily in hypergraphs, and methods to dynamically learn the underlying topology from data (ICLR, 2024; ICML, 2024; TMLR, 2025). These works provide both theoretical insights and practical tools for modeling complex systems where interactions go beyond pairwise relationships.

[5] Interpretable AI systems for science

As AI systems are increasingly deployed in scientific and high-stakes domains, ensuring their interpretability and transparency has become a fundamental requirement. My work in this area focuses on developing models and methods that allow practitioners to understand, validate, and control the behavior of neural networks, particularly in complex and data-intensive settings. A distinctive aspect of this research is its grounding in real-world scientific applications, where interpretability is not only desirable but necessary to support domain experts. In this context, we developed interpretable models for analyzing neural activity (Journal of Neural Engineering, 2024), identifying early indicators of neurological disorders from handwriting (Movement Disorders Clinical Practice, 2025), and studying the origin of archaeological materials such as Levantine ceramics (Machine Learning: Science and Technology, 2025). Together, these contributions highlight how interpretability can act as a bridge between machine learning and scientific discovery, enabling AI systems that are not only accurate, but also trustworthy and actionable in multidisciplinary contexts.

[6] AI Models for High-Energy Physics Experiments

In high-energy physics (HEP), machine learning models are increasingly used to analyze large volumes of data generated by particle collisions, for example to distinguish rare signals from background events in real time (*triggering*). In this context, interpretability is particularly important, as models must be validated against established physical theories and integrated into complex experimental pipelines. Building on the research direction outlined in [5], we developed interpretable neural models for muon detection in collaboration with the ATLAS groups at the University of Rome and the University of Liverpool (Computing and Software for Big Science, 2023), as well as graph-based representations of simulated collision data to study potential signatures of new particles, including supersymmetric candidates (Scientific Reports, 2025).

[7] AI for Energy Systems and Infrastructure

The integration of renewable energy sources introduces significant challenges in the management and reliability of modern power systems, due to the inherent variability and uncertainty of generation. In this context, my work has explored the use of AI models, particularly graph-based approaches, to improve forecasting and decision-making in energy networks. Building on the methodological developments in graph and geometric deep learning [4] and interpretable systems [5], we developed models for forecasting solar and wind power generation that explicitly account for spatial and relational dependencies across the grid (IEEE Transactions on Industry Applications, 2025; Expert Systems with Applications, 2025; Applied Energy, 2023). A key focus of this work is not only predictive accuracy, but also robustness and interpretability, which are essential for deployment in operational settings.

[8] AI for Earth Observation and Remote Sensing

Earth observation systems generate large volumes of heterogeneous data, including satellite imagery and multispectral measurements, which are essential for monitoring environmental changes and managing natural resources. In this context, my work has focused on developing AI models that can effectively leverage these data while reducing their reliance on large annotated datasets. In particular, we investigated self-supervised learning techniques for remote sensing data and methods for 3D change detection in complex environments (IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023; European Journal of Remote Sensing, 2025). These approaches enable the extraction of meaningful representations from limited supervision, improving both scalability and applicability in real-world scenarios. More broadly, this line of research contributes to the development of AI tools for large-scale environmental monitoring, supporting applications in climate analysis, land use assessment, and infrastructure management.

Peer recognition

Editorial roles: I contribute to the evolution of research in machine learning as Area Chair for two of the most prestigious conferences in the field (NeurIPS and ICLR), and as Action Editor for *Transactions of Machine Learning Research (TMLR)*. In 2021, I also served as chair of the Statistical Pattern Recognition Techniques Technical Committee of the International Association for Pattern Recognition.

Fellowships: I am currently an elected Junior Fellow at the Sapienza School for Advanced Studies (2024-2026), within the academic class of science and technology, and an affiliate researcher at the Istituto Nazionale di Fisica Nucleare (INFN), reflecting my involvement in interdisciplinary research at the intersection of AI, physics, and engineering.

Research projects: My research activities are supported by competitive funding and leadership roles in national and international projects. I currently act as Principal Investigator for a Sapienza grant (*CENTS: A Cost-adaptive Neural Network framework for Semantic Communication*), local PI for a PRIN project (*CLASSICS*), and Work Package leader for the European project 6G-GOALS and for the Norwegian Research Council project *RELAY*.

PhD supervision: I am involved in the supervision and mentoring of early-stage researchers, currently supervising 9 PhD students (with 9 additional PhD students having recently graduated), including several positions funded by external institutions (e.g., Translated, Pi School, Leonardo, Fastweb).

Awards: In recognition of my teaching and outreach activities, I was awarded a Google Faculty Award in 2020 to support initiatives on machine learning education, diversity, and inclusion.

Invited talks: I have also been invited to deliver talks and keynotes in a variety of academic and industrial contexts, including the University of Cambridge, Università della Svizzera italiana (Lugano), TIM Playmaker, Huawei Cloud Workshop, the National PhD Programme in Technologies for Fundamental Research in Physics and Astrophysics, the Italian National Institute of Health, and the national congress of SISET.

ADDITIONAL INFORMATION

Other contributions to the research community

Beyond my core research activities, I have been actively involved in fostering connections between academic research, industry, and society, with a focus on the dissemination and practical adoption of AI technologies.

I have served as scientific supervisor for several industrial research projects, leading to both scientific outputs and technology proof-of-concepts. These include a multi-year collaboration on audio analysis systems with the Consorzio Interuniversitario Nazionale per l'Informatica, a project with the Parco Archeologico del Colosseo on data-driven approaches to environmental monitoring and weed management, and a recent collaboration with Leonardo on the robustness of AI models, focusing on adversarial attacks and system reliability.

I have also been engaged in outreach and dissemination activities aimed at promoting AI literacy across different sectors of society. Before the COVID-19 pandemic, I co-founded and led the no-profit association Italian Association for Machine Learning, organizing seminars, workshops, and public events across Italy. More recently, I have coordinated a Sapienza-wide initiative on AI literacy for teaching staff, developed within the Digilab research center, and I currently coordinate the industrial liaison program of the Sapienza Master's Degree in Data Science, organizing yearly events involving 25–30 companies and institutions. In addition, I regularly contribute to outreach through invited talks in diverse contexts, including schools, industry events, and public-facing initiatives.

As part of my commitment to education and knowledge dissemination, I authored the book *Alice's Adventures in a Differentiable Wonderland*, which provides an accessible introduction to the design of neural networks. The book has reached a wide audience, with approximately 4,000 copies sold worldwide and over 500,000

Template CV Soci Accademia di Ingegneria e Tecnologia

downloads from the associated website. This work reflects my broader interest in making advanced machine learning concepts accessible to a wider community of students, researchers, and practitioners.