

## Summary

- Principal Research Scientist with the AI Innovation Team at **Red Hat** and Research Affiliate at **MIT MechE**, working at the intersection of AI and engineering.
- Specializing in Generative AI, Generative Optimization, and Probabilistic Modeling, with a specific focus on **Inference-Time Scaling**, **Test-Time Adaptation**, **Vision-Language Alignment**, and **Few-Shot Generation**.
- Leading research on Probabilistic Inference to develop efficient, grounded Foundation Models tailored for data-constrained engineering domains.

## Experience

### Principal Research Scientist, Red Hat

Boston, Massachusetts, USA  
*June 2025 - Present*

- AI Innovation Team
  - Probabilistic Inference for Vision and Language Models
  - Inference-Time Scaling and Reasoning for LLMs
  - **its-hub** Development and vLLM Gateway Integration
  - Context Optimization for AgentOps

### Research Affiliate, Massachusetts Institute of Technology

Cambridge, Massachusetts, USA  
*Jan 2026 - Present*

- DeCoDE Lab. Department of Mechanical Engineering
  - Inference-Time Scaling for Constrained Generative Design
  - Iterative Self-Training for CAD Program Synthesis

### Applied Scientist, Amazon

Seattle, Washington, USA  
*April 2024 - June 2025*

- Home Innovation and GenAI Team
  - Grounded Vision-Language Models
  - Evaluation for Text-to-Image Models (CVPR)
  - Detection and Ranking Algorithms for Amazon Visual Shopping
  - Subject-Driven Generative Models for AI Creative Studio

### Visiting Researcher, UCL Centre for Artificial Intelligence

London, UK  
*Jan 2024 - March 2024*

- Host: David Barber
  - Multi-Resolution Convolutional Models for Long Sequences (NeurIPS)
  - Bayesian Inference for Language Models

### PhD Internships and Collaborations

- Researcher (PhD Intern), Microsoft Research Cambridge, MA, USA, *Jun 2023 - Sept 2023*
  - ML and Statistics Group. Hosts: David Alvarez Melis, Nicolo Fusi
  - Dynamic Vocabulary Augmentation for LLMs
- Research Collaborator, MIT-IBM AI Lab Cambridge, MA, USA, *Jan 2023 - June 2023*
  - Model Alignment Team. Host: Akash Srivastava
  - Generative Models for Systems with Constraints (NeurIPS)
  - Aligning Language Models with Negative Data

- Specialized Language Models for Enterprise Domains
- Research Scientist (PhD Intern), IBM Research Zurich, Switzerland, *Jun 2022 - Nov 2022*
  - Accelerated Discovery Team. Hosts: Matteo Manica, Teodoro Laino
  - Multitask Language Models for Text and Chemistry (ICML)
  - Open-source library GT4SD for conditional generative models
- Applied Scientist (PhD Intern), Amazon Science Cambridge & London, UK, *Jul 2021 - Oct 2021*
  - Alexa Team. Hosts: Yunlong Jiao, Emine Yilmaz
  - Domain Agnostic Subpopulation Generalisation

#### **Research Engineer, NNAISENSE**

Lugano, Switzerland  
*Jan 2019 - Jan 2020*

- Deep Learning Team. Managers: Christian Osendorfer, Jonathan Masci
  - Structured Latent Variable Models
  - NeuralODE Algorithms for High-Range Event Camera Streams

#### **Co-Founder, SecretAIry (formerly GAiA)**

Rome, Italy  
*July 2017 - Jan 2019*

- Chatbots to enhance Workplace Communication
  - Selected among 100+ startups to join the EnLabs Incubator

## **Education**

#### **PhD, Generative Machine Learning**

Technical University of Denmark, Lyngby, Denmark  
*June 2020 - Dec 2023*

- Few-Shot Generative Models (ICML)
- Multitask Language Models for Conditional Molecule Generation (ICML)
- Diffusion Models for Generative Engineering Design and Topology Optimization (NeurIPS)
- Thesis: Learning Generative Models with Limited Data
  - Supervisor: Ole Winther; Co-supervisor: Søren Hauberg

#### **Visiting PhD Student, MIT School of Engineering**

Cambridge, Massachusetts, USA  
*Jan 2023 - Sept 2023*

- Constrained Diffusion Models for Engineering Design (NeurIPS & Patent)
- Improving Generative Constraint Satisfaction using Invalid Designs (TMLR)
- Evaluating Vision-Language Models for Engineering Tasks (Journal)
- Research on LLM Agents for CAD design. Co-developer of `text2cad`.
  - Host: Faez Ahmed, DeCoDE Lab
- Selected for Post-Doctoral and Research Scientist roles at MIT and MIT-IBM (declined).

#### **Master's Degree, Data Science**

Sapienza University, Rome, Italy  
*Sept 2016 - Nov 2018*

- Excellence Path & Summa Cum Laude
- Thesis: Multimodal Learning for Scene Understanding
  - Supervisor: Aris Anagnostopoulos; External Supervisor: Boris Chidlovskii

#### **Visiting Graduate Student, NYU Tandon School of Engineering**

NYC, New York, USA  
*Sept 2017 - Jan 2018*

- Visualization and Data Analytics Research Center. Host: Enrico Bertini
  - Built an interactive entity retrieval tool to investigate 10M documents

#### **Master's Degree, Mechanical Engineering**

Sapienza University, Rome, Italy  
*Sept 2014 - Jan 2017*

- Summa Cum Laude
- Thesis: Bubble Dynamics in Turbulent Shear Flows
  - Supervisor: Carlo Massimo Casciola; Co-supervisor: Paolo Gualtieri

#### **Bachelor's Degree, Mechanical Engineering**

Sapienza University, Rome, Italy  
*Sept 2009 - May 2014*

- Thesis: Rapid Prototyping of Metallic Manufacturing

## Selected Publications & Patents

|  |  |
|--|--|
| <b>Bootstrapping Image-to-CAD Program Synthesis via Geometric Feedback</b><br><u>GIANNONE</u> , DORIS, NOBARI, XU, SRIVASTAVA, AHMED                                   | under-review<br>2026                       |
| <b>Mitigating Premature Exploitation in Particle-based Monte Carlo for ITS</b><br><u>GIANNONE</u> , XU, NAYAK, AWHAD, SUDALAIRAJ, XU, SRIVASTAVA                       | under-review<br>2025                       |
| <b>Generative optimization models for machine learning</b><br><u>GIANNONE</u> , SRIVASTAVA, AHMED  | US Patent (MIT & IBM)<br>2025              |
| <b>Feedback-Driven Vision-Language Alignment</b><br><u>GIANNONE</u> , LI, FENG, PEREVODCHIKOV, CHEN, MARTINEZ  | under-review<br>2025                       |
| <b>Be More Specific: Evaluating Object-centric Realism in Synthetic Images</b><br>LIANG, CORNEANU, FENG, <u>GIANNONE</u> , MARTINEZ                                    | CVPR<br>2025                               |
| <b>Evaluating Vision-Language Models for Engineering Design</b> Springer Artificial Intelligence Review<br>PICARD, EDWARDS, DORIS, MANN, <u>GIANNONE</u> , ALAM, AHMED | <br>2025                                   |
| <b>NITO: Neural Implicit Fields for Resolution-free Topology Optimization</b><br>NOBARI, REGENWETTER, <u>GIANNONE</u> , AHMED  | TMLR<br>2025                               |
| <b>Reparameterized Multi-Resolution Convolutions for Long Sequence Modelling</b><br>CUNNINGHAM, <u>GIANNONE</u> , ZHANG, DEISENROTH                                    | NeurIPS<br>2024                            |
| <b>Constraining Generative Models for Engineering Design with Negative Data</b><br>REGENWETTER, <u>GIANNONE</u> , SRIVASTAVA, GUTFREUND, AHMED                         | TMLR<br>2024                               |
| <b>Aligning Optimization Trajectories with Diffusion Models</b><br><u>GIANNONE</u> , SRIVASTAVA, WINTHER, AHMED  | NeurIPS<br>2023                            |
| <b>Diffusing the Optimal Topology: A Generative Optimization Perspective</b><br><u>GIANNONE</u> , AHMED  | IDETC23<br>2023                            |
| <b>Unifying Molecular and Textual Representations via Multi-task LM</b><br>CHRISTOFIDELLIS*, <u>GIANNONE</u> *, BORN, WINTHER, LAINO, MANICA                           | ICML<br>2023                               |
| <b>Accelerating Material Design with GT4SD</b><br><i>GT4SD Team (Core Contributor)</i>   | Nature npj Computational Materials<br>2023 |
| <b>Few-Shot Diffusion Models</b><br><u>GIANNONE</u> , NIELSEN, WINTHER   | SBM@NeurIPS<br>2022                        |
| <b>SCHA-VAE: Hierarchical Context Aggregation for Few-Shot Generation</b><br><u>GIANNONE</u> , WINTHER   | ICML<br>2022                               |
| <b>Method and apparatus for semantic segmentation and depth completion</b><br>CHIDLOVSKII, <u>GIANNONE</u>   | US Patent (NAVER)<br>2022                  |
| <b>JM1: Worst-group Generalization by Group Interpolation</b><br><u>GIANNONE</u> , HAVRYLOV, MASSIAH, YILMAZ, JIAO   | NeurIPS-W<br>2021                          |
| <b>Hierarchical Few-Shot Generative Models</b><br><u>GIANNONE</u> , WINTHER  | NeurIPS-W<br>2021                          |
| <b>Transformation-aware Variational Autoencoders</b><br><u>GIANNONE</u> , SAREMI, MASCI, OSENDORFER  | Technical Report<br>2020                   |
| <b>Input-filtering NeuralODEs for spiking data</b><br><u>GIANNONE</u> , ANOOSHEH, QUAGLINO, D'ORO, MASCI, GALLIERI   | NeurIPS-W<br>2020                          |
| <b><math>\mathcal{T}</math>-VAE: No Representation without Transformation</b><br><u>GIANNONE</u> , MASCI, OSENDORFER   | NeurIPS-W<br>2019                          |
| <b>Learning Common Representation from RGB and Depth Images</b><br><u>GIANNONE</u> , CHIDLOVSKII   | CVPR-W<br>2019                             |

## Projects & Open Source

|  |      |
|--|------|
| its-hub: A Python library for inference-time scaling | 2025 |
|--|------|

- Contributor.
- Inference-Time Scaling for Language Models.
- Focus on Mathematical Reasoning.
- Contributed Entropic Particle Filtering algorithms and new benchmark.

#### **Text2CAD: Democratizing Engineering Design. Prompt by Prompt.**

2023

- Co-Lead.
- DesignX. Team of engineers and researchers based at MIT and Caltech.
- Generative tool that allows users to create CAD models using natural language prompts.
- The tool is designed to be user-friendly and accessible to non-experts, enabling a wide range of users to quickly create complex CAD models without the need for specialized training.

#### **GT4SD: Generative Toolkit for Scientific Discovery**

2022

- Core Contributor.
- Library leveraging conditional generative models for accelerated discovery.
- Work on Diffusion Models for images and 3D molecule conformation. The GFlowNet framework. Property Prediction module. Public Hub for model upload. Training Pipelines. Documentation. Tutorials. Testing. CI/CD. Server and Client API. Docker Images for CPU and GPU.

### **Grants & Awards**

#### **GPU Grant, LUMI-G, EuroHPC**

PI, Efficient Pre-training of Large Generative Models for Constrained Design

Copenhagen, Denmark

*Nov 2023*

#### **Grant, Otto Mønsted's Foundation**

Research Grant

Copenhagen, Denmark

*Dec 2022*

#### **Grant, Independent Research Fund Denmark**

DFF PhD Grant

Lyngby, Denmark

*Jun 2020*

#### **Grant, Perception as Generative Reasoning**

Awarded Complimentary Conference Registration by DeepMind

NeurIPS 2019

*Oct 2019*

#### **Grant, Pi School**

Full Tuition for the School of AI (3% acceptance rate)

Rome, Italy

*Oct 2018*

#### **Certificate of Award, Tsinghua University**

Prize for Outstanding Accomplishments, Deep Learning Summer School

Beijing, China

*Aug 2018*

#### **1st Pick, Excellence Path, Master's Degree, Data Science**

Admission based on the First year's Academic Performance

Rome, Italy

*Mar 2018*

Participation in the School for Advanced Studies

### **Academic Service**

#### **Reviewer**

Conference: ICML19, ICCV19, AAAI20, ICML21 (top 10%), AISTATS21, ICML22, NeurIPS22, CVPR23, NeurIPS23, ICML24, ICLR25, CVPR25, NeurIPS25, ICLR26, ICML26

Journal: TPAMI, TMLR

Workshop: NeurIPS-IBW20, NeurIPS-MetaLearn21, ICML-DeployableGenAI23, ACL-LanguageMolecules24

#### **Teaching**

Teaching: Deep Learning (DTU 02456), Bayesian Machine Learning (DTU 02477), Advanced Machine Learning (DTU 02460)

Supervision: two special courses (9 months), two master's thesis (6+6 months), 18 final projects

### **Talks**

Algorithmic Methods for Data Mining (Sapienza University), Bayesian Reading Group (DTU), MLLS Center (KU), UCL-NLP (UCL), Amazon Alexa (Cambridge), DeCoDE Lab (MIT)

### **Volunteering**

PAISS18, NeurIPS18, ELLIS Unit Copenhagen, MLLS

## **Skills**

### **Languages**

- Python (proficient); R, Matlab (good knowledge); C, Java, JavaScript (basic knowledge)

### **Research**

- Accelerate, HF Transformers, LaTeX, NLTK, OpenCV, PyTorch, SpaCy, TensorFlow, verl

### **Software**

- AWS, CVX, Docker/podman, FastAPI, Git, GitHub Actions, Gradio, Linux, MinIO, MongoDB, MySQL, Travis, vLLM, LangGraph, LangFlow, Langfuse, Cline, Cursor, OpenRouter