# Federico Vaggi, PhD

**Seattle, United States**

vaggi.federico@gmail.com

**github** ▪ **google scholar** ▪ **linkedin**

## PROFESSIONAL SUMMARY

I'm a senior staff scientist with 10+ years of experience shaping ML strategy for confidential, high-impact projects at Alphabet and Amazon. Since 2016, I've focused on applied ML research, developing new methods with impact measured through products rather than publications. I've built research programs that grew into revenue-generating products, co-led design of ML platforms adopted across organizations, and provided technical insights that shaped billion-dollar business decisions. My expertise spans generative AI, causal inference, optimization, differential equations and Bayesian methods, and I have a knack for matching the right solution to the problem at hand.

## EXPERIENCE

**ML Lead/Senior Staff Scientist**      *Google X (Computational Biology Project)*      **Jan 2021-present**

- **Pioneered LLMs for biological discovery** by designing a multimodal transformer architecture that integrated diverse biological data sources across organisms. **Achieved 30% improvement** in prediction accuracy while scaling to 1000x larger training datasets, enabling the team to tackle projects in data-scarce organisms that were previously infeasible.
- **Drove org-wide ML transformation** to hybrid and LLM-based approaches, mentoring senior bioinformaticians to enable adoption; this strategy became the foundation for the org's two largest revenue-generating projects.
- **Built a hybrid simulation platform** combining stiff differential equations with neural differential equations, solving complex technical challenges including stiff adjoints and steady-state optimization, fully integrated into JAX for differentiable scientific computing.
- **Solved a core data-integration bottleneck** by developing a novel hierarchical Bayesian Expectation-Maximization algorithm to merge heterogeneous biological datasets (across organisms, sites, and conditions), expanding usable data 3x for the largest partner project.
- **Developed uncertainty-aware active learning pipeline** leveraging parametric mechanistic models and low-rank MCMC, **reducing experimental costs by up to 60%** by optimally selecting which experiments to run.

**Principal Applied Scientist**      *Amazon Seller Partner Services*      **June 2019- Jan 2021**

- **Established causal inference standards across a 5000+ person organization:** Led development of scalable observational methods that became the org-wide framework for evaluating billion-dollar fee changes and seller programs.
- **Lead work on seller profitability estimation:** Developed a new neural network regressor architecture (involving transfer learning, low data finetuning, domain adaptation) whose outputs featured in Jeff Bezos's annual shareholder letters.
- **Co-architected ML Pigeon,** the next-generation internal ML platform adopted organization-wide with **100+ production projects,** cutting ML deployment costs and time by **80%.**

**Senior Applied Scientist**      *Amazon Fulfillment by Amazon*      **April 2017-June 2019**

- **Pioneered Bayesian synthetic control methods** for causal impact estimation, outperforming existing CausalImpact baselines and enabling reliable intervention analysis across multiple business dimensions.
- **Generated $50M in annual savings** through a novel multi-dimensional mixture density network that accurately estimated weight/dimensions of third-party seller items.
- **Built the science organization:** grew FBA's science team from 2 to 15 scientists, mentored juniors into senior roles, and served as bar raiser for Amazon's technical interviews, ensuring hiring quality across the org.

**Researcher**      *INRIA-Ecole Normale Supérieure*      **February 2016-March 2017**

- **Developed novel causal GAN architecture and applied it to biological imaging:** First to apply generative adversarial networks to fluorescent microscopy, publishing breakthrough ICCV paper and open-source implementation

adopted by the research community.

– **Developed scalable ML solutions for industry:** Built structured regularization frameworks (lasso variants, trend filtering) enabling efficient training of very large-scale linear models in collaboration with AXA, a major European insurance company.

**Researcher**                     *Fondazione Edmund Mach*                     **January 2013-January 2016**

– **Led cross-domain method development:** Developed novel computational biology approaches and successfully adapted them to solve problems in other domains, demonstrating broad applicability of technical innovations.
– **Created viral algorithmic solution:** Designed graph theory-based conference networking system that gained international attention, featured in Harvard Business Review and adopted at multiple scientific conferences worldwide.

## EDUCATION

**Postdoctoral Researcher**    Microsoft Research - University of Trento, COSBI                     **2013**
**PhD, Systems Biology**    IFOM-IEO Institute, University of Milan                     **2010**
**MSc, Bioinformatics**    University of Milan, Bicocca                     **2006**
**BSc, Physics**    Imperial College London                     **2004**

## SKILLS

– **Generative AI & Differentiable Programming**: Large language models, multimodal transformers, neural ODEs, JAX/PyTorch, data-efficient fine-tuning.
– **Causal Inference & Decision Science**: Observational frameworks at internet scale, synthetic controls, instrumental variables, experimental design for billion-dollar policy changes.
– **Platform & Ecosystem Building**: Org-wide ML platforms (100+ production projects), standards for causal inference across 5000+ scientists/analysts, strategy-setting for multi-year ML roadmaps.
– **Statistical Modeling & Optimization**: Bayesian and frequentist inference, uncertainty quantification, non-linear and large-scale optimization, Gaussian processes, ensemble methods.

## SELECTED PUBLICATIONS AND PATENTS

– **30+ peer-reviewed publications** in computational biology, ML, and computer vision (720+ citations, h-index 11).
– **US Patent US20250218545**: Negative Data Augmentation for Training LLMs - novel method for low-data generative model training exploiting pseudo-likelihood in logits
– **ICCV 2017**: GANs for Biological Image Synthesis (155 citations) - first application of generative models to fluorescent microscopy, introducing cross-channel GAN architecture