Marc Harary

Email: marc.harary@yale.edu
Mobile: +1 609-751-6524
Github: github.com/marc-harary

Education

- **Massachusetts Institute of Technology** — Cambridge, Massachusetts
  Non-degree Coursework — *2023 - Present*
  Courses: *Optimization Methods*

- **Yale College** — New Haven, Connecticut
  B.A. Mathematics and Philosophy — *2017 - 2022*
  Courses: *Data Structures, Algorithms, Artificial Intelligence, Bayesian Statistics, Dynamical Systems, General Chemistry, Neurobiology*

Publications

- **PathoGen: Immunohistochemistry *in silico* via Cascaded Latent Diffusion for Semantic Synthesis**, M Harary, W Lotter, EM Van Allen. *(In Submission)*, 2024.
- **Fluoroformer: Scaling multiple instance learning to multiplexed images via attention-based channel fusion**, M Harary, EM Van Allen, W Lotter. *(In Submission)*, 2024.
- **SHIELD: Secure Haplotype Imputation Employing Local Differential Privacy**, M Harary. *arXiv preprint arXiv:2309.07305*, 2024.
- **Efficient Algorithms for the Sensitivities of the Pearson Correlation Coefficient and Its Statistical Significance to Online Data**, M Harary. *arXiv preprint arXiv:2405.14686*, 2024.
- **Kirigami: Large Convolutional Kernels Improve Deep Learning-Based RNA Secondary Structure Prediction**, M Harary, C Zhang, AM Pyle. *arXiv preprint arXiv:2406.02381*, 2024.
- **3D Printed Quadcopters**, LS Dai, MA Harary, CT Pompei, SL Shan, MJ Tu. *New Jersey Governor's School of Engineering and Technology*, 2016.

Skills and Training Summary

- **Languages**: Python, C/C++/CUDA, LaTeX, LISP, Julia, MATLAB, R, Bash
- **Frameworks**: PyTorch, TensorFlow, Keras, Numpy, Matplotlib, Scipy, Sympy, Pandas
- **Areas of computer science**: Theoretical computer science, computational biology & bioinformatics, deep learning & artificial intelligence, image processing & computer vision, cryptography
- **Areas of empirical research**: Genomics, computational RNA biology, proteomics, cognitive neuroscience, deep learning
- **Misc.**: Proof-based algorithm design, manuscript and essay writing

Experience

- **Van Allen Lab** — Dana-Farber Cancer Institute
  Computational Biologist — *September 2023 - Present*
  - Personally developed CycleGAN model to perform neural style transfer between hematoxylin and eosin (H&E) and immunohistochemistry (IHC) images
  - Personally wrote conditional Generative Adversarial Network (cGAN) for same task
  - Conceived of and wrote a denoising diffusion probabilistic model (DDPM) called PathoGen
  - First author on manuscript "PathoGen: Immunohistochemistry *in silico* via Cascaded Latent Diffusion for Semantic Synthesis" in submission to Conference on Computer Vision and Pattern Recognition (CVPR)

- **Lotter Lab** — Dana-Farber Cancer Institute
  Computational Biologist — *September 2023 - Present*
  - Wrote several PyTorch pipelines for attention-based multiple instance learning (ABMIL) for survival analysis and tissue classification of hematoxylin and eosin (H&E) slides for non-small cell lung cancer (NSCLC) samples
  - Wrote overall survival (OS) prediction model for multiplexed immunofluorescence (mIF) samples
  - Invented the Fluoroformer architecture, a Transformer-like model designed to perform explainable multimodal fusion of mIF images
  - First author on "Fluoroformer: Scaling multiple instance learning to multiplexed images via attention-based channel fusion," submitted to ML4Health conference

- **Cho Lab** — Broad Institute of MIT and Harvard
  Computational Biologist — *September 2022 - Present*
  - Member of Cho Lab, which specializes in cryptographic algorithms for genomic analysis

- Collaborated to develop secure methods for genome-wide association studies (GWAS) using techniques from differential privacy
- Wrote encrypted genomic imputation algorithm via the Li-Stephens hidden Markov model of ancestral recombination

- **Pyle Lab** Yale University
  Computational Biologist *September 2020 - Present*
  - Personally wrote Kirigami, a state-of-the-art deep learning-based RNA secondary structure prediction pipeline
  - Pipeline used in the Critical Assessment of Structure of Protein (CASP) competition
  - First author publication expected early 2023

- **ProteoWise, Inc.**
  Algorithm Engineer *May 2021 - August 2022*
  - Conceived novel molecular kinetics-based segmentation algorithm for Western blot images, increasing sensitivity to picomolar concentrations and signal-to-noise ratio (SNR) by an order of magnitude
  - Derived curve-fitting subroutines from first principles from optimization theory
  - Implemented algorithms from scratch in CUDA/C++
  - Received equity in company for contributions

- **Cognitive and Neural Computation Lab** Yale University
  Research Assistant *March 2020 - September 2021*
  - Engineered computer graphics subroutines in C++ for mathematical models of human visual cognition
  - Wrote generative modeling scripts in Julia for computational physics

- **Turk-Browne Lab** Yale University
  Research Assistant *October 2018 - September 2019*
  - Conceived fMRI experiment to observe pattern completion in the hippocampus
  - Awarded human subjects to perform associated behavioral experiments
  - Conducted data analysis in R and MATLAB

- **Social Robotics Laboratory** Yale University
  Research Assistant *March 2018 - October 2018*
  - Programmed robot to facilitate lab meetings in Java

- **Clinical & Affective Neuroscience Laboratory** Yale University
  Research Assistant *March 2018 - August 2018*
  - Assisted with meta-analysis for neuroimaging study

Marc Harary

A little-known fact in computer science is that the first program run by the Von Neumann architecture was written by a mathematical biologist to simulate digital life. By pursuing a PhD in Computer Science at Columbia, my aspiration would be to contribute to among the greatest challenges of this century for computer scientists, namely algorithmic privacy for vulnerable data like genomic records. My extensive experience in pioneering my own novel projects across many of the world's premier research facilities, keen ability to innovate in many of the interstices of computer science and biology, and predisposition for mathematical elegance all contribute to the strength of my candidacy.

Having faced rejection in the last admissions cycle, I have since significantly bolstered my research profile, accomplished significant technical achievements, and, most importantly, collaborated extensively with Gamze Gursoy in Columbia's Computer Science department. She has now offered to be my academic advisor for the upcoming spring term, having expressed significant interest in my independence and mathematical abilities as a computer science researcher. In light of the close alignment of her lab's interests and my own, I contend that I stand to make significant contributions to Columbia's research community and to receive highly beneficial mentorship from an environment uniquely suited to my passions.

Following my previous admissions cycle, I became even more fiercely determined to strengthen my research abilities, resolute in my goal to join a computer science program, and ambitious in presenting a stronger application this cycle. At the Broad Institute and the New York Genome Center, I worked with Gamze Gursoy to make significant strides towards solving a **significant open question in genomic security** by applying differential privacy (DP) to genotype imputation by developing **SHIELD (Secure Haplotype Imputation Employing Local Differential Privacy)** (https://arxiv.org/pdf/2309.07305). Drawing on my **skills both as a**

mathematician and bioinformatician, I wrote **several pages of complex mathematical proofs** of SHIELD's correctness and optimality based on induction across genotype markers. We now have a **highly polished software package** that is soon to be in submission to **RECOMB** with myself as **first author**.

In addition, I have worked on several other projects that highlight the **wide, interdisciplinary scope** of my academic portfolio and my abilities as a **mathematical scientist**. I developed the **quadratic sweep algorithm** (https://github.com/marc-harary/QuadraticSweep**),** a potentially **groundbreaking result in combinatorial geometry**. Its purpose is answer the simple following question: given a dataset consisting of n points in the plane, which subset of size k has the highest coefficient of determination or Pearson correlation coefficient (PCC)? In a **first-author paper** that I **intend to submit to SIAM**, I showed that this combinatorial optimization problem is solvable in efficient time by borrowing advanced concepts from robust statistics, combinatorics, computational geometry, and matroid theory. Conducting further mathematical research on the PCC, I discovered **key properties of correlation analysis** in a **solo-author** manuscript (https://arxiv.org/abs/2405.14686) submitted to **NeurIPs,** accompanied by a highly polished computational statistics package in C and Cython (https://github.com/marc-harary/sensitivity). Critically, these results can be applied directly to enforcing privacy for genome-wide association studies (GWAS), where the **local sensitivity** of the PCC is vital for determining the exposure of subjects to deanonymization attacks.

A **highly versatile and adaptable** researcher, I made substantial contributions to AI and computational histopathology (CPath) at the Dana-Farber Cancer Institute (DFCI) since my rejection last fall. In order to scale attention-based multiple instance learning (ABMIL) to multiplexed immunofluorescence (mIF) images, **I invented the Fluoroformer architecture**, a

cutting-edge Transformer-like attention mechanism that performs multimodal fusion. With Bill Lotter's leadership, we achieved a concordance index (C-index) of over 80% for survival analysis on a large non-small cell lung cancer (NSCLC) cohort and submitted a paper to **ML4Health** on which I am **first author**. In doing so, I provided ample evidence of my ability to rapidly **self-teach new topics** and **identify gaps in the literature**. Under the guidance of Eliezer Van Allen, I demonstrated **creative ambition** and **interdisciplinary collaboration** when I conceived of **PathoGen, a denoising diffusion probabilistic model (DDPM)** to simulate immunohistochemistry (IHC) *in silico*. I am now writing a **first-author** manuscript for **CVPR** in the coming weeks. With the support of Gamze Gursoy, my goal would be to continue similar interdepartmental relationships to reach similar heights in privacy research.

Under Gamze Gursoy's guidance, our objective is to spend my tenure at Columbia, if I were admitted, contributing to privacy projects in bioinformatics of equal importance. Beyond imputation, I have a strong interest in **applying my work on the PCC to GWAS**, another foundational pillar of contemporary bioinformatics. Applying my background in deep learning, I also hope to **develop DDPMs with DP codebooks** for biological applications, preserving the privacy of donors to these critical reference panels. A final interest of mine is **electronic health records (EHRs)**, which I believe stand as an excellent application of privacy, natural language processing (NLP), and deep learning for computational biology.

With an avid passion for computation, mathematics, and bioinformatics, my objective in following a PhD in computer science from a renowned institution like Columbia would be to pursue a career as a principal investigator (PI) in academia. In a world increasingly dominated by systems whose obscure structure remains impenetrable to researchers, computer science is, as a field, in urgent need of researchers willing to scrutinize AI and algorithms in general with a high

degree of mathematical rigor. The mounting ethical challenges posed by systems with the potential to infringe upon subject privacy necessitate a new generation of researchers, one with the ambition of designing novel protocols that are able to reap the benefits of the ongoing AI renaissance without causing harm to the broader society. Nowhere are privacy dilemmas more pressing or complex than in their intersection with bioethics and bioinformatics, where the most precious of our data stand to transform modern medicine but also expose our most intimate secrets. Under the guidance of Gamze Gursoy, I hope to nurture my abilities as a mathematician and computer scientist in order to secure a bright and safe future for computer science in the twenty-first century.