

# Yair Schiff

[yairschiff@cs.cornell.edu](mailto:yairschiff@cs.cornell.edu)

[yair-schiff.github.io](https://yair-schiff.github.io)

## RESEARCH INTERESTS

---

Generative modeling, Optimal transport, AI for Science, AI for Social Good



## EDUCATION

---

**CORNELL UNIVERSITY** – DEPARTMENT OF COMPUTER SCIENCE, New York, NY

2022 – Present

*PhD student in Computer Science*

*Advisor: Professor Volodymyr Kuleshov*

- *Fellowships:* Hal & Inge Marcus PhD Fellowship (2022-2023)
- *Awards:* Cornell Tech Outstanding TA Award (2023)
- *Teaching:* Applied Machine Learning (TA Fall 2022), Deep Learning (TA Spring 2023)

**NEW YORK UNIVERSITY** – COURANT INSTITUTE OF MATHEMATICAL SCIENCES, New York, NY

May 2019

*MS in Computer Science, GPA: 3.97/4.00*

- *Relevant Coursework:* Advanced Machine Learning, Artificial Intelligence, Computer Vision, Deep Learning, Graphical Processing Units, Machine Learning, Mathematics of Deep Learning, Predictive Analytics

**UNIVERSITY OF PENNSYLVANIA** – COLLEGE OF ARTS AND SCIENCES, Philadelphia, PA

May 2014

*BA Summa Cum Laude with Distinction in Economics, GPA: 3.93/4.00*

- *Academic Honors:* Phi Beta Kappa, Dean's list 2010-2014



## RESEARCH AND PUBLICATIONS

---

### PUBLICATIONS

Caduceus: Bi-Directional Equivariant Long-Range

[ICML 2024](#)

DNA Sequence Modeling

**Yair Schiff**, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, Volodymyr Kuleshov

DySLIM: Dynamics Stable Learning by Invariant Measure for Chaotic Systems

[ICML 2024](#)

**Yair Schiff**, Zhong Yi Wan, Jeffrey B. Parker, Stephan Hoyer, Volodymyr Kuleshov,

Fei Sha, Leonardo Zepeda-Núñez

InfoDiffusion: Representation Learning

[ICML 2023](#)

Using Information Maximizing Diffusion Models

Yingheng Wang, **Yair Schiff**, Aaron Gokaslan, Weishen Pan, Fei Wang,

Christopher De Sa, Volodymyr Kuleshov

Semi-Autoregressive Energy Flows: Exploring Likelihood-Free

[ICML 2023](#)

Training of Normalizing Flows

Phillip Si, Zeyi Chen, Subham Sekhar Sahoo, **Yair Schiff**, Volodymyr Kuleshov

Learning with Stochastic Orders

[ICLR 2023](#)

Carles Domingo-Enrich, **Yair Schiff**, Youssef Mroueh

(Notable Top 25% acceptance)

Semi-Parametric Inducing Point Networks and Neural Processes

[ICLR 2023](#)

Richa Rastogi, **Yair Schiff**, Alon Hachohen, Zhaozhi Li, Ian Lee, Yuntian Deng,

Mert R. Sabuncu, Volodymyr Kuleshov

Cloud-Based Real-Time Molecular Screening Platform  
with MolFormer

Brian Belgodere\*, Vijil Chenthamarakshan\*, Payel Das\*, Pierre Dognin\*,  
Toby Kurien\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*,  
Jarret Ross\*, **Yair Schiff**\*, Richard A. Young\*

[ECML PKDD 2022 Demo Track](#)

(\*alphabetical order  
equal contribution)

Optimizing Functionals on the Space of Probabilities with  
Input Convex Neural Networks

David Alvarez-Melis, **Yair Schiff**, Youssef Mroueh

[Transactions of Machine Learning Research](#)

Augmenting Molecular Deep Generative Models with  
Topological Data Analysis Representations

**Yair Schiff**\*, Vijil Chenthamarakshan\*, Samuel Hoffman\*,  
Karthikeyan Natesan Ramamurthy\*, Payel Das\*

[ICASSP 2022](#)

(\*equal contribution)

Predicting Deep Neural Network Generalization with Perturbation Response Curves  
**Yair Schiff**, Brian Quanz, Payel Das, Pin-Yu Chen

[NeurIPS 2021](#)

Tabular Transformers for Modeling Multivariate Time Series

Inkit Padhi, **Yair Schiff**, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin,  
Jarret Ross, Ravi Nair, Erik Altman

[ICASSP 2021](#)

Image Captioning as an Assistive Technology: Lessons Learned from  
VizWiz 2020 Challenge

Pierre Dognin\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*,  
Jarret Ross\*, **Yair Schiff**\*, Richard Young, Brian Belgodere

[Journal of AI Research](#)

(\*alphabetical order  
equal contribution)

## WORKSHOPS

Simple and Effective Masked Diffusion Language Models

Subham Sekhar Sahoo, Marianne Arriola, **Yair Schiff**, Aaron Gokaslan,  
Edgar Marroquin, Justin T Chiu, Alexander Rush, Volodymyr Kuleshov

[ICML Workshop 2024](#)

Advancing DNA Language Models: The Genomics Long-Range Benchmark

Evan Trop, Chia-Hsiang Kao, Mckinley Polen, **Yair Schiff**, Bernardo P. de Almeida,  
Aaron Gokaslan, Thomas Pierrot, Volodymyr Kuleshov

[AAAI Workshop 2024](#)

[ICLR Workshop 2024](#)

Optimizing Functionals on the Space of Probabilities with  
Input Convex Neural Networks

David Alvarez-Melis, **Yair Schiff**, Youssef Mroueh

[NeurIPS Workshop 2021](#)

Spotlight presentation

Gi and Pal Scores: Deep Neural Network Generalization Statistics

**Yair Schiff**, Brian Quanz, Payel Das, Pin-Yu Chen

[ICLR Workshop 2021](#)

Characterizing the Latent Space of Molecular Deep Generative Models  
with Persistent Homology Metrics

**Yair Schiff**, Vijil Chenthamarakshan, Karthikeyan Natesan Ramamurthy, Payel Das

[NeurIPS Workshop 2020](#)

Spotlight presentation

Alleviating Noisy Data in Image Captioning with Cooperative Distillation

Pierre Dognin\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*,  
Jarret Ross\*, **Yair Schiff**\*

[CVPR Workshop 2020](#)

(\*alphabetical order,  
equal contribution)

## PREPRINTS

Cross-species plant genomes modeling at single nucleotide resolution using a pre-trained DNA language model

[bioRxiv](#)

Jingjing Zhai, Aaron Gokaslan, **Yair Schiff**, Ana Berthel, Zong-Yan Liu, Zachary R Miller  
Armin Scheben, Michelle C Stitzer, Cinta Romy, Edward S. Buckler, Volodymyr Kuleshov

Auditing and Generating Synthetic Data with Controllable Trust Trade-offs

[arXiv](#)

Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh,  
Aleksandra Mojsilovic, Jiri Navartil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti,  
Jerret Ross, **Yair Schiff**, Radhika Vedpathak, Richard A. Young

## TALKS AND PRESENTATIONS

Topological Data Analysis and Beyond Workshop

[NeurIPS 2020](#)

- Presented spotlight poster “Characterizing the Latent Space of Molecular Deep Generative Models” ([video](#))

VizWiz Grand Challenge Workshop

[CVPR 2020](#)

- Presented winning submission to VizWiz Grand Challenge ([video](#))
- Presented “Alleviating Noisy Data in Image Captioning with Cooperative Distillation” ([video](#))

## OPEN-SOURCE CONTRIBUTIONS

Caduceus

[Github.com](#)

- Authored codebase for long-range DNA sequence modeling using newly proposed Caduceus model.

swirl-dynamics

[Github.com](#)

- Added a project for modeling dynamical systems using a regularized objective that aims to preserve systems’ invariant measures

SPIN: Semi-Parametric Inducing Point Networks and Neural Process

[Github.com](#)

- Implemented Inducing Point Neural Processes and wrote code for training and evaluation

TabFormer: Tabular Transformers for Modeling Multivariate Time Series

[Github.com](#)

- Wrote code for training and evaluating GPT-like models on tabular data to generate new, synthetic data that matches the underlying distributions of the real table variables

pytorch-PPUU: Prediction and Policy-learning Under Uncertainty

[Github.com](#)

- Added a new dataset on which the self-driving policy could be trained
- Enhanced the self-driving vehicle’s policy to enable dynamic lane changes

## PROFESSIONAL SERVICES

- ICML 2024 reviewer
- NeurIPS 2023 reviewer – [Top Reviewer Recipient](#)



## WORK EXPERIENCE

**GOOGLE**, New York, NY

May 2023 – Oct 2023

*Student Researcher*

- Researched new methods for stabilizing autoregressive rollouts of dynamical system models
- Contributed to internal Google and open-source libraries for modeling dynamical systems

**IBM WATSON MACHINE LEARNING**, New York, NY

Aug 2019 – Aug 2022

*Cognitive Software Developer*

- Contributed to continuous development and testing of Watson Machine Learning products
- Facilitated weekly Cloud releases, Cloud Pak for Data platform releases, and the launch of AutoAI feature engineering on relational data, AutoAI Time Series, AutoAI Notebooks, and Federated Learning products
- Published 6 medium.com articles about Watson Machine Learning product releases
- Received Outstanding Technical Achievement Award for work on the release of AutoAI feature engineering on relational data
- Received CrushIT Team Excellence Award as part of the Watson Machine Learning Training team

*Research Contributor to IBM Research AI Challenges*

Aug 2019 – Aug 2022

- Volunteered to contribute to IBM Research AI challenges, working with the Trusted AI Department
- Member of the first-place winning team in the 2020 VizWiz Grand Challenge: Image Captioning as an Assistive Technology for the Visually Impaired
- Co-authored with the Trusted AI team on several publications in the fields of Generative Modeling, Molecular Discovery, Deep Learning Generalization, and AI for Social Good
- Received two 2021 IBM Research Accomplishments awards for contributions to (1) trustworthy AI generative modeling and (2) deployment of large-scale transformer models on OpenShift environments

**SIMON-KUCHER AND PARTNERS**, New York, NY

Sept 2014 – Aug 2017

*Consultant*

- Advised global companies spanning various industries – including internet, media, consumer electronic goods, and chemicals – on areas for better revenue capture
- Synthesized large data sets (e.g., 300 million+ client transactions), customer research (surveys & conjoint studies sent to thousands of respondents), and secondary research to create solutions to client needs



## GROUPS AND AFFILIATIONS

---

**GRADS FOR GENDER INCLUSION IN COMPUTING**, New York, NY

Present

*Member*

- Member of a group dedicated to combating harassment, pushing for policy change, and creating supportive spaces

**STUDENT-APPLICANT SUPPORT PROGRAM**, New York, NY

Present

*Volunteer*

- Provide feedback and support to students applying to graduate programs in Computer Science and related fields

**PHD PEER MENTOR PROGRAM**, New York, NY

Present

*Mentor*

- Meet with fellow PhD student to provide guidance about graduate school life at Cornell Tech



## SKILLS

---

- *Programming Languages:* C++, Java, Python
- *Deep Learning Frameworks:* PyTorch, PyTorch Lightning
- *Data Tools:* Excel, Stata, Tableau
- *Foreign Languages:* Fluent in Hebrew