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, NY2022 – 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, NYMay 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, PAMay 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**

Simple and Effective Masked Diffusion Language Models [NeurIPS 2024](https://arxiv.org/abs/2406.07524)

*Subham Sekhar Sahoo, Marianne Arriola,* ***Yair Schiff****, Aaron Gokaslan,*

*Edgar Marroquin, Justin T Chiu, Alexander Rush, Volodymyr Kuleshov*

Auditing and Generating Synthetic Data with Controllable Trust Trade-offs [IEEE JETCAS](https://arxiv.org/abs/2304.10819)

*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*

Caduceus: Bi-Directional Equivariant Long-Range [ICML 2024](https://arxiv.org/abs/2403.03234)

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](https://arxiv.org/abs/2402.04467)

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

*Fei Sha, Leonardo Zepeda-Núñez*

InfoDiffusion: Representation Learning [ICML 2023](https://arxiv.org/abs/2306.08757)

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](https://openreview.net/forum?id=Skrk3StS2g)

Training of Normalizing Flows

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

Learning with Stochastic Orders [ICLR 2023](https://openreview.net/forum?id=P3PJokAqGW)

*Carles Domingo-Enrich,* ***Yair Schiff****, Youssef Mroueh* *(Notable Top 25% acceptance)*

Semi-Parametric Inducing Point Networks and Neural Processes [ICLR 2023](https://openreview.net/forum?id=FE99-fDrWd5)

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

*Mert R. Sabuncu, Volodymyr Kuleshov*

Cloud-Based Real-Time Molecular Screening Platform [ECML PKDD 2022 Demo Track](https://arxiv.org/abs/2208.06665)

with MolFormer *(\*alphabetical order*

*Brian Belgodere\*, Vijil Chenthamarakshan\*, Payel Das\*, Pierre Dognin\*, equal contribution)*

*Toby Kurien\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*,*

*Jarret Ross\*,* ***Yair Schiff\*****, Richard A. Young\**

Optimizing Functionals on the Space of Probabilities with [Transactions of Machine Learning Research](https://arxiv.org/abs/2106.00774)

Input Convex Neural Networks

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

Augmenting Molecular Deep Generative Models with [ICASSP 2022](https://arxiv.org/abs/2106.04464)

Topological Data Analysis Representations *(\*equal contribution)*

***Yair Schiff\*****, Vijil Chenthamarakshan\*, Samuel Hoffman\*,*

*Karthikeyan Natesan Ramamurthy\*, Payel Das\**

Predicting Deep Neural Network Generalization with Perturbation Response Curves [NeurIPS 2021](https://arxiv.org/abs/2106.04765)

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

Tabular Transformers for Modeling Multivariate Time Series [ICASSP 2021](https://ieeexplore.ieee.org/document/9414142)

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

*Jarret Ross, Ravi Nair, Erik Altman*

Image Captioning as an Assistive Technology: Lessons Learned from [Journal of AI Research](https://arxiv.org/abs/2012.11696)

VizWiz 2020 Challenge *(\*alphabetical order*

*Pierre Dognin\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*, equal contribution)*

*Jarret Ross\*,* ***Yair Schiff****\*, Richard Young, Brian Belgodere*

**Workshops**

Advancing DNA Language Models: The Genomics Long-Range Benchmark [AAAI Workshop 2024](https://llms4science-community.github.io/papers/LLMs4Bio24_paper_12.pdf)

*Evan Trop, Chia-Hsiang Kao, Mckinley Polen,* ***Yair Schiff****, Bernardo P. de Almeida,* [ICLR Workshop 2024](https://mlgenx.github.io/)

*Aaron Gokaslan, Thomas Pierrot, Volodymyr Kuleshov*

Optimizing Functionals on the Space of Probabilities with [NeurIPS Workshop 2021](https://arxiv.org/abs/2106.00774)

Input Convex Neural Networks *Spotlight presentation*

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

Gi and Pal Scores: Deep Neural Network Generalization Statistics [ICLR Workshop 2021](https://arxiv.org/abs/2104.03469)

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

Characterizing the Latent Space of Molecular Deep Generative Models [NeurIPS Workshop 2020](https://arxiv.org/abs/2010.08548)

with Persistent Homology Metrics *Spotlight presentation*

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

Alleviating Noisy Data in Image Captioning with Cooperative Distillation [CVPR Workshop 2020](https://arxiv.org/abs/2012.11691)

*Pierre Dognin\*, Igor Melnyk\*, Youssef Mroueh\*, Inkit Padhi\*, Mattia Rigotti\*, (\*alphabetical order,*

*Jarret Ross\*,* ***Yair Schiff****\* equal contribution)*

**Preprints**

Cross-species plant genomes modeling at single nucleotide resolution [bioRxiv](https://www.biorxiv.org/content/10.1101/2024.06.04.596709v1)

using a pre-trained DNA language model

*Jingjing Zhai, Aaron Gokaslan,* ***Yair Schiff****, Ana Berthel, Zong-Yan Liu, Zachary R Miller*

*Armin Scheben, Michelle C Stitzer, Cinta Romay, Edward S. Buckler, Volodymyr Kuleshov*

**Talks and Presentations**

Topological Data Analysis and Beyond Workshop [NeurIPS 2020](https://tda-in-ml.github.io/papers)

* Presented spotlight poster “Characterizing the Latent Space of Molecular Deep Generative Models” ([video](https://slideslive.com/38941579/characterizing-the-latent-space-of-molecular-deep-generative-models-with-persistent-homology-metrics?ref=account-folder-62106-folders))

VizWiz Grand Challenge Workshop [CVPR 2020](https://vizwiz.org/workshops/2020-workshop/)

* Presented winning submission to VizWiz Grand Challenge ([video](https://ivc.ischool.utexas.edu/~yz9244/VizWiz_workshop/videos/MMTeam-oral.mp4))
* Presented “Alleviating Noisy Data in Image Captioning with Cooperative Distillation” ([video](https://ivc.ischool.utexas.edu/~yz9244/VizWiz_workshop/videos/103-poster.mp4))

**Open-Source Contributions**

Caduceus [Github.com](https://github.com/kuleshov-group/caduceus)

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

swirl-dynamics [Github.com](https://github.com/google-research/swirl-dynamics)

* 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](https://github.com/RichRast/SPIN)

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

TabFormer: Tabular Transformers for Modeling Multivariate Time Series [Github.com](https://github.com/IBM/TabFormer)

* 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](https://github.com/Atcold/pytorch-PPUU)

* 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**

* ICLR 2025 reviewer
* ICML 2024 reviewer
* NeurIPS 2023 reviewer – [Top Reviewer Recipient](https://nips.cc/Conferences/2023/ProgramCommittee#top-reivewers)

# WORK EXPERINCE

**Instadeep**, New York, NY May 2024 – *Present*

*PhD Researcher Intern*

* + - * Applying generative modeling techniques to genomic sequences
      * Investigating control mechanisms for guided sequence generation

**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 Challenge*s 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