

Chao Peter Yang

Researcher in Graph ML, Symbolic Music AI, and Agentic Systems
+1-773-901-0080 chaopeter.yang@gmail.com [linkedin.com/in/chao-peter-yang](https://www.linkedin.com/in/chao-peter-yang)
site-cpyang.github.io github.com/cpyang123

EDUCATION

Duke University

Aug 2024 – May 2026

- ◆ *M.S. in Interdisciplinary Data Science*, **MIDS Merit Scholarship (70%)**

GPA: 3.93/4.0

- ◆ Coursework: NLP, Statistical Modeling, Data Engineering, ML Theory & Algorithms, Probabilistic ML, Generative AI.

University of Michigan – Ann Arbor

Aug 2018 – May 2021

- ◆ *B.S. in Data Science and Mathematics*
- ◆ Honors: **Highest Honors in Data Science (one of two)**, **University Honors** (2019, 2021).

RESEARCH EXPERIENCE

Interpretable Machine Learning Lab, Duke University

Aug 2024 – Present

Research Assistant (Advised by Prof. Cynthia Rudin & Stephen Ni-Hahn)

Durham, NC

- ◆ Authored and built **ProGress**—structured symbolic music generation via rule-guided **Discrete Diffusion**; 45-subject study beat baselines on quality with only ~3M params; *under review, NeurIPS 2025 Creative Track*.
- ◆ Researched and developed a custom implementation of **DiffPool** for **Heterogeneous GNN** used in musical analysis in **PyTorch**, reducing validation **cross-entropy loss** by 60% with additional **hyperparameter tuning**.
- ◆ Developed **Proximal Policy Optimization** for **Graphical Neural Network** to enable **Reinforcement Learning from Human Feedback**, enabling automated, personalized musical analysis. *Under Review ACMHCI '26*

University of Michigan, Honors Student Researcher

Sep 2020 – May 2021

- ◆ Developed new music classification methods using Musical Instrument Digital Interface (MIDI) and **LSTM neural networks** resulting in 82% **5-fold-cross-validated** accuracy, more than 10% improvement over conventional **ML methods**.
- ◆ Improved models using **supervised machine learning methods** like **Support Vector Machines**, **Decision Trees**, **Ensemble Methods**, **K-nearest neighbors** etc..
- ◆ Advisors: Prof. Edward L. Ionides (Statistics), Prof. Daniel Forger (Mathematics).

University of Michigan, Student Researcher

Jan 2020 – May 2020

- ◆ Utilized **Fast Fourier Transform (FFT)** and **convolution reverb** techniques to simulate acoustic properties of various venues, revealing significant environmental effects on sound perception.
- ◆ Successfully authored and **secured research funding** from Prof. Joseph Gascho, managing financial transactions between vendors and the Stearns Collection at the University of Michigan.
- ◆ Authored and secured project funding; supervised by Prof. Mark E. Newman (Physics).

PUBLICATIONS

- ◆ Stephen Ni-Hahn*, **Chao Péter Yang***, Mingchen Ma, Cynthia Rudin, Simon Mak, Yue Jiang. *ProGress: Structured Music Generation via Graph Diffusion and Hierarchical Music Analysis*. Submitted to the Thirty-Ninth Conference on Neural Information Processing Systems (NeurIPS 2025), Creative AI Track. *Under review*.
- ◆ Stephen Ni-Hahn, Jerry Zhu, **Chao Péter Yang**, et al. *SchenkerLink: Human-in-the-Loop Hierarchical Music Analysis with Uncertainty-Aware Graph Link Prediction*. Submitted to the ACM CHI Conference on Human Factors in Computing Systems (CHI '26) / PACM on HCI, 2026. *Under review*.
- ◆ Undergraduate Honors Thesis: *The Classical-Romantic Dichotomy: A Machine Learning Approach*
https://ionides.github.io/students/cpyang_honors_thesis.pdf

SERVICES

- ◆ Reviewer: NeurIPS 2025 AI4Music Workshop

SELECTED INDUSTRY EXPERIENCE

Amazon — Amazon Robotics, Data Scientist Intern

May 2025 – Aug 2025

- ◆ Designed and deployed a **root-cause investigation agent** integrating logs/telemetry/tickets; **reduced troubleshooting from days to 2.5 minutes at 75% hit rate**.
- ◆ Built a **unified agentic framework** on **LangGraph** + **Amazon Bedrock** to standardize development and evaluation of internal agents.
- ◆ Created a **production evaluation harness** (*LLM-as-a-Judge* + **Langfuse**) enabling reproducible, large-scale A/B and regression testing.

Informa PLC/ Curinos, Senior Data Science Analyst, Modeling

October 2023 – June 2024

- ◆ Researched and developed industry-level **nonlinear** elasticity models for **Asset-Liability Management (ALM)** to

predict acquisition and other portfolio balances for regional banks and credit unions, resulting in improved prediction vs. legacy models in terms of out-of-sample R^2 .

- ◆ Created automated ad-hoc regression notebooks with **PySpark** for creating, testing, and validating models with different configurations, reducing the time to build proof-of-concept models by half.

Informa PLC/ Curinos, *Data Science Analyst II*

April 2022 – October 2023

- ◆ Led ML engineering team to migrate legacy modeling pipeline from using **Cloudera** to **Databricks**, coordinating across teams to schedule testing, promotion, and release plans, leading to more than \$100k in annual savings for platform expenses, and a 30% decrease in data processing time on average. **Acknowledged in company-wide town hall meeting**
- ◆ Tuned **nonlinear hierarchical price elasticity models** en masse for multiple major US banks, each with 10,000+ model segments, resulting in improved fit in terms of both AIC and R^2 with a significantly higher rate of convergence.
- ◆ Set up and automated custom **SQL** procedures to clean, wrangle, and transform client's data feed to be used in the modeling pipeline, eliminating the need for manual model data refreshes.

Informa PLC/ Curinos, *Data Science Analyst*

Aug 2021 – April 2022

- ◆ Converted local, single-threaded, legacy modeling pipeline to use **SparkR** and **Cloudera**, reducing run time for model fitting by up to 30 times.
- ◆ Performed **Exploratory Data Analysis (EDA)** for client banks to tune and reconfigure their models and data segments, leading to better performing **price elasticity models** in terms of MAPE, R^2 , and rate of convergence.
- ◆ Installed and managed more than 10,000 **price elasticity models** per client bank to predict and optimize their deposit portfolio across a wide range of interest rates, with precise **Model Risk Management** documentation.

SELECTED PROJECTS

ProfMatch (Graph-based RAG) [Project link](#)

Fall 2024

- ◆ Built an LLM-driven professor recommendation system using **LightRAG** over a graph of faculty/research topics; implemented a **Streamlit** UI with NetworkX visualizations.

SanAssist (Healthcare Data Analytics Platform) [Repo](#)

Fall 2024

- ◆ Fine-tuned **GPT-2 (LoRA)** on a domain corpus (val perplexity 3.32); deployed with **Docker** to **AWS ECR** + **App Runner**; load-tested to **10k concurrent VUs**.
- ◆ Built ETL (Databricks, Pandas) and a metrics API (**Squirrels**) with CI/CD via GitHub Actions.

Muscribe (Music Transcription)

Fall 2023

- ◆ Trained **Transformer** and **CRNN** models for audio→symbolic transcription; reported note-level precision/recall against open baselines with reduced training resources.

AWARDS AND HONORS

- ◆ *Duke University*: MIDS Scholarship (70%)
- ◆ *University of Michigan*: Highest Honors in Data Science
- ◆ *University of Michigan*: University Honors (2019, 2021)
- ◆ *Rados Deszö Violin Competition*: Gold Medalist
- ◆ *Central & Eastern European Schools Association (CEESA)*: Silver in Tennis Doubles
- ◆ *Danube Valley Athletic Conference (DVAC)*: Gold in Tennis Doubles

CERTIFICATIONS

- ◆ *DeepLearning.AI*: Neural Networks and Deep Learning
- ◆ *DeepLearning.AI*: Structuring Machine Learning Projects
- ◆ *DeepLearning.AI*: Improving Deep Neural Networks: Hyper-parameter Tuning, Regularization and Optimization
- ◆ *DeepLearning.AI*: Convolutional Neural Network
- ◆ *DeepLearning.AI*: Generative AI with Large Language Models
- ◆ *Google*: Share Data Through the Art of Visualization
- ◆ *DataCamp*: Introduction to Scala

TECHNICAL SKILLS

Programming	Python (PySpark, PyTorch, sklearn, SciPy Pandas, Numpy), R (SparkR, dplyr), SQL, Rust, Scala, C++, Javascript, MATLAB, Git
Machine Learning	Regressions, Boosted Trees, RLHF, Deep Learning, LSTM, CNN, GNN, Transformers, Diffusion
Gen AI/Agents	Langgraph, ReAct Agents, Amazon Bedrock, Langfuse, RAG, MCP Server
Big Data/Databases	Spark, Databricks, AWS, SQL Server, PostgreSQL, MongoDB, SQLite
Visualization	matplotlib, ggplot, Tableau, Excel Dashboard
Languages	Fluent: English, Hungarian, Mandarin Chinese — Conversational: German, Japanese

ACTIVITIES & SERVICE

- ◆ Volunteering: Greater Chicago Food Depository (2023); Budapest Festival Orchestra (2017); Habitat for Humanity (2015)
- ◆ Music/Tennis: University Orchestra (2018); Kroó György Community Orchestra (2015–2018); UMich Club Tennis (2019–2020)

LANGUAGES

English (Professional); Hungarian (Native); Mandarin Chinese (Native); German (Intermediate); Japanese (Conversational)

REFERENCES

Prof. Edward L. Ionides, Associate Chair for Undergraduate Studies and Professor

Department of Statistics, University of Michigan
ionides@umich.edu, +1 (734) 615-3332

Prof. Mark E. Newman, Anatol Rapoport Distinguished University Professor of Physics

Department of Physics, University of Michigan
mejn@umich.edu, +1 (734) 764-4437

Ryan Schulz, Client Success & Modeling Director

Modeling Team, Curinos, Inc.
ryan.schulz@curinos.com, +1 (609) 558-0713