

Minh-Anh To

minhanh.to@duke.edu – (919) 475-8485 – [LinkedIn](#) – [Portfolio](#)

EDUCATION

Duke University - The Graduate School

Aug 2023 – Exp. May 2025

M.S. - Statistical Science - GPA: 3.91/4.0

Durham, NC

Full Tuition Scholarship; Vingroup Scholar; American Statistical Association Member (ASA).

Hanoi National University of Education

Aug 2014 – May 2018

B.S. - Mathematics (Honors) - GPA: 3.65/4.0 (Top 5%)

Hanoi, Vietnam

National Mathematics Scholarship (Top 1% Applicants); Vice President of Student Council; Dean's List.

SKILLS

Programming: Python (NumPy, pandas, scikit-learn, PyTorch, TensorFlow), R (tidyverse, Shiny, dplyr, tidymodels), SQL, Tableau, Power BI, Git, HTML, Cloud Services.

AI/ML: Generative AI (LangGraph, Agentic AI), Deep Learning (CNNs, RNNs), Machine Learning (XGBoost, Random Forest, Logistic Regression, Support Vector Machine, K-means, KNN), Natural Language Processing (NLP).

Statistics: Hypothesis Testing, Predictive Modeling, Causal Inference, Bayesian Modeling, Hierarchical Modeling.

WORK EXPERIENCE

ChessTutorAI

Aug 2024 – Present

AI Engineer

Remote, NC

- Developed an **AI Chess Coach** using **Generative AI** and expertise from a **chess grandmaster** to enhance player skill acquisition and strategic decision-making.
- Designed an **agentic AI workflow** using **LangChain** and **LangGraph** to retrieve chess concepts from an expert-curated dataset, increasing model **retrieval accuracy** by **36%**.
- Optimized Large Language Model (LLM) responses through **few-shot learning** and **chain-of-thought prompting**, minimizing hallucinations and achieving over **99% semantic similarity accuracy**.
- Integrated an **interactive** LLM-based grading module into the AI Chess Coach product, providing **personalized** feedback that enhanced **user experience** and increased **engagement** by **25%**.

Duke University School of Medicine

May 2024 – Aug 2024

Data Science Intern

Durham, NC

- Transformed **10 years** of unstructured biomedical data into a structured database (5K+ mRNA sequences) using **SQL** and **Pandas**, enabling **scalable predictive modeling** for antibody production.
- Collaborated with a **cross-functional** team to engineer 35 mRNA structural features by devising statistical formulas and implementing custom Python functions, improving model performance by **75%**.
- Developed, fine-tuned, and validated tree-based models (XGBoost, Random Forest)** with **scikit-learn**, achieving **90%** predictive accuracy and reducing runtime by **48%**.
- Conducted **feature importance analysis** to identify five key mRNA properties that improved yield predictions, driving cost savings of **tens of thousands of dollars**.
- Authored a technical report on model performance and business impact, selected from 50 candidates to present to **200+ attendees** at the Duke CFAR Retreat 2024.

Vietnam Academy of Science and Technology

Jun 2022 – Aug 2023

Assistant Researcher

Hanoi, Vietnam

- Proved the practicality of the **bootstrap method** for goodness-of-fit testing using estimator properties; presented findings to **25 researchers** and secured a significant grant from VinIF Vietnam.
- Led a weekly Bayesian Statistics seminar for **30+ students** on **simulation & optimization**, including live **R** coding.

PROJECT EXPERIENCE

[Nuclei Detection in Breast Cancer Images](#) (Python, Deep Learning, CNNs): Integrated a breast cancer image dataset with 125K+ labeled nuclei into HuggingFace for efficient access, conducted EDA for nuclei distributions across hospitals, and fine-tuned a YOLOv8 model in PyTorch, achieving 85% accuracy in cancerous nuclei detection.

[Hyperparameter Optimization Shiny App](#) (R, Machine Learning, Shiny): Developed an interactive Shiny app to compare tuning methods (Grid Search, Random Search, Bayesian Optimization), delivering real-time insights on model performance. Integrated LASSO, Ridge, and Random Forest, boosting prediction accuracy by 24%.

[Using Generalized Linear Models for Sparrow Dataset](#) (R, Inferential Statistics): Analyzed a 19-year sparrow dataset using hierarchical GLMs with autoregressive terms to model time correlation, reducing BIC by 27%.