Quan Nguyen

Department of Computer Science
Princeton University
Princeton, NJ 08540

Phone +1 (765) 721-5818
Email qn9088@princeton.edu
Website krisnguyen135.github.io

Introduction

My research lies at the interaction of probabilistic machine learning, decision theory, and scientific discovery. I am broadly motivated by the following question: *How can we design algorithms that make intelligent decisions under uncertainty, in a way that is both theoretically principled and practically impactful for real-world scientific and engineering settings?*

Much of my work has focused on active learning and Bayesian experimental design, where the goal is to identify which experiment or data point to collect next in order to best achieve a specified goal. These problems arise in hyperparameter tuning of machine learning models, drug discovery, materials design, and more broadly in any domain where experiments are expensive and data are scarce. The unifying theme of my research has been to (1) develop budget-aware, nonmyopic algorithms that reason about the long-term consequences of decisions, and (2) introduce new measures of uncertainty that better capture scientific objectives.

While much of my work is motivated by challenges in science and engineering, the core contributions are broadly applicable to machine learning: designing decision-making algorithms that balance exploration and exploitation, quantifying uncertainty in complex domains, and developing scalable information-theoretic objectives.

Research accomplishments

Nonmyopic active learning and search

A central challenge in experimental design is that greedy, one-step-ahead strategies can be highly suboptimal. In my Ph.D. work, I developed algorithms for nonmyopic decision-making that approximate the optimal long-horizon policy.

- In Nguyen et al. [7], we developed multifidelity active search algorithms that exploit cheap, low-fidelity data to accelerate discovery at reduced cost.
- In Nguyen and Garnett [6], we introduced a framework for multiclass search with diminishing returns, demonstrating how to capture discovery diversity while remaining sample-efficient.
- In Nguyen et al. [8], we proposed a local Bayesian optimization method that maximizes the probability of descent, improving robustness in high-dimensional search spaces.

Together, these contributions show how decision-theoretic approaches can overcome the limitations of greedy strategies, especially in high-cost scenarios where labeling is expensive.

Experimental design for scientific discovery

I have collaborated extensively with chemists, physicists, and materials scientists to demonstrate the utility of machine learning-guided experiments in domain applications.

- With collaborators at Colorado School of Mines and Princeton, I developed methods for predicting metastable polymorph synthesizability and probabilistic material stability, integrating uncertainty estimates directly into experimental workflows [9, 10].
- In Liu, Nguyen et al. [3], we showed how diversity-driven active learning accelerates the designs of MOFs (metal–organic frameworks) for ammonia adsorption, yielding important environmental applications.

These projects highlight my ability to not only develop new methodology but also validate it in interdisciplinary, real-world domains.

Information-theoretic measures of uncertainty

A recurring theme in my postdoctoral research is the role of information gain in guiding experiments. Classical information-theoretic measures such as mutual information often fail in high-dimensional or structure domains.

- In recent works [4, 5], I introduced Vendi information gain (VIG), a generalization of mutual information that captures both informativeness and diversity of outcomes.
- We showed that VIG provides better guidance in experimental design and active learning tasks, as well as in scenarios where the aim is to balance exploitation with diverse exploration.

This line of work opens the door to a broader family of uncertainty measures tailored to discovery, and it forms the foundation of my future research program.

Future research agenda

My future work will build upon past accomplishments to advance the theory and practice of experimental design. I envision three main directions.

New information-theoretic objectives

I will develop principled objectives for guiding data acquisition that go beyond classical measures like mutual information. My recent work on Vendi information gain (VIG) demonstrates how diversity and informativeness can be captured simultaneously. Future work will strengthen its theoretical foundations (e.g., submodularity [1], regret bounds [11]), develop scalable estimators using variance reduction and importance sampling [2], and extend these ideas to structured domains such as molecules and graphs. VIG's flexibility promises a unified approach across many experimental settings such as multiobjective, multifidelity, and preference-based designs.

Budget-aware policy design

Building on my Ph.D. work in nonmyopic planning, I aim to develop policies that optimize entire sequences of queries under budget constraints, rather than relying on one-step heuristics. Inspired by sequence training in large language models, this includes learning query policies from simulated trajectories, using transformer-based models to capture long-horizon dependencies, and training with sequence-level rewards that reflect overall utility. These methods will yield resource-aware policies that align directly with task-specific goals. An immediate step will be extending my earlier deep imitation learning work to train query policies that optimize entire sequences under budgets.

Applications of experimental design

Finally, I will continue to validate and expand experimental design methodology across diverse domains. In science and engineering, this includes ongoing research on materials discovery, drug design, and sustainable chemistry with collaborators at Princeton and Mines. Beyond the natural sciences, I plan to extend these ideas to the social sciences such as political science and psychology, where data are costly and uncertainty quantification is essential. By grounding theory in real-world problems, this agenda will demonstrate experimental design as a general framework for data-efficient decision-making.

Broader impact and vision

My long-term vision is to develop machine learning systems that accelerate the pace of discovery by making principled, resource-aware decisions about where to look next. The combination of probabilistic modeling, decision theory, and interdisciplinary applications uniquely positions my work to contribute both fundamental advances in machine learning and practical impact in the sciences.

In parallel, I will continue to build collaborations across disciplines, target machine learning venues as well as domain journals, and mentor students in developing the next generation of tools at the intersection of machine learning and science.

References

[1] Daniel Golovin and Andreas Krause. Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

[2] Alireza Haghighat and John C Wagner. Monte Carlo variance reduction with deterministic importance functions. *Progress in Nuclear Energy*, 42(1):25–53, 2003.

[3] Tsung-Wei Liu, Quan Nguyen, Adji Bousso Dieng, and Diego A. Gómez-Gualdrón. Diversity-driven, efficient exploration of a MOF design space to optimize MOF properties. *Chemical Science*, 15(45):18903–18919, 2024.

[4] Quan Nguyen and Adji Bousso Dieng. Quality-Weighted Vendi Scores And Their Application To Diverse Experimental Design. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[5] Quan Nguyen and Adji Bousso Dieng. Vendi information gain: An alternative to mutual information for science and machine learning. *arXiv preprint*, 2025. arXiv:2505.09007.

[6] Quan Nguyen and Roman Garnett. Nonmyopic Multiclass Active Search with Diminishing Returns for Diverse Discovery . In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[7] Quan Nguyen, Arghavan Modiri, and Roman Garnett. Nonmyopic Multifidelity Active Search. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[8] Quan Nguyen, Kaiwen Wu, Jacob R. Gardner, and Roman Garnett. Local Bayesian optimization via maximizing descent probability. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[9] Andrew Novick, Diana Cai, Quan Nguyen, Roman Garnett, Ryan Adams, and Eric Toberer. Probabilistic prediction of material stability: integrating convex hulls into active learning. *Materials Horizons*, 11(21):5381–5393, 2024.

[10] Andrew Novick, Quan Nguyen, Matthew Jankousky, M.B̃rooks Tellekamp, Eric S. Toberer, and Vladan Stevanović. Basin-Size Mapping: Prediction of Metastable Polymorph Synthesizability Across TaC–TaN Alloys. *Journal of the American Chemical Society*, 147(5):4419–4429, 2025.

[11] Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. *Operations Research*, 66(1):230–252, 2018.